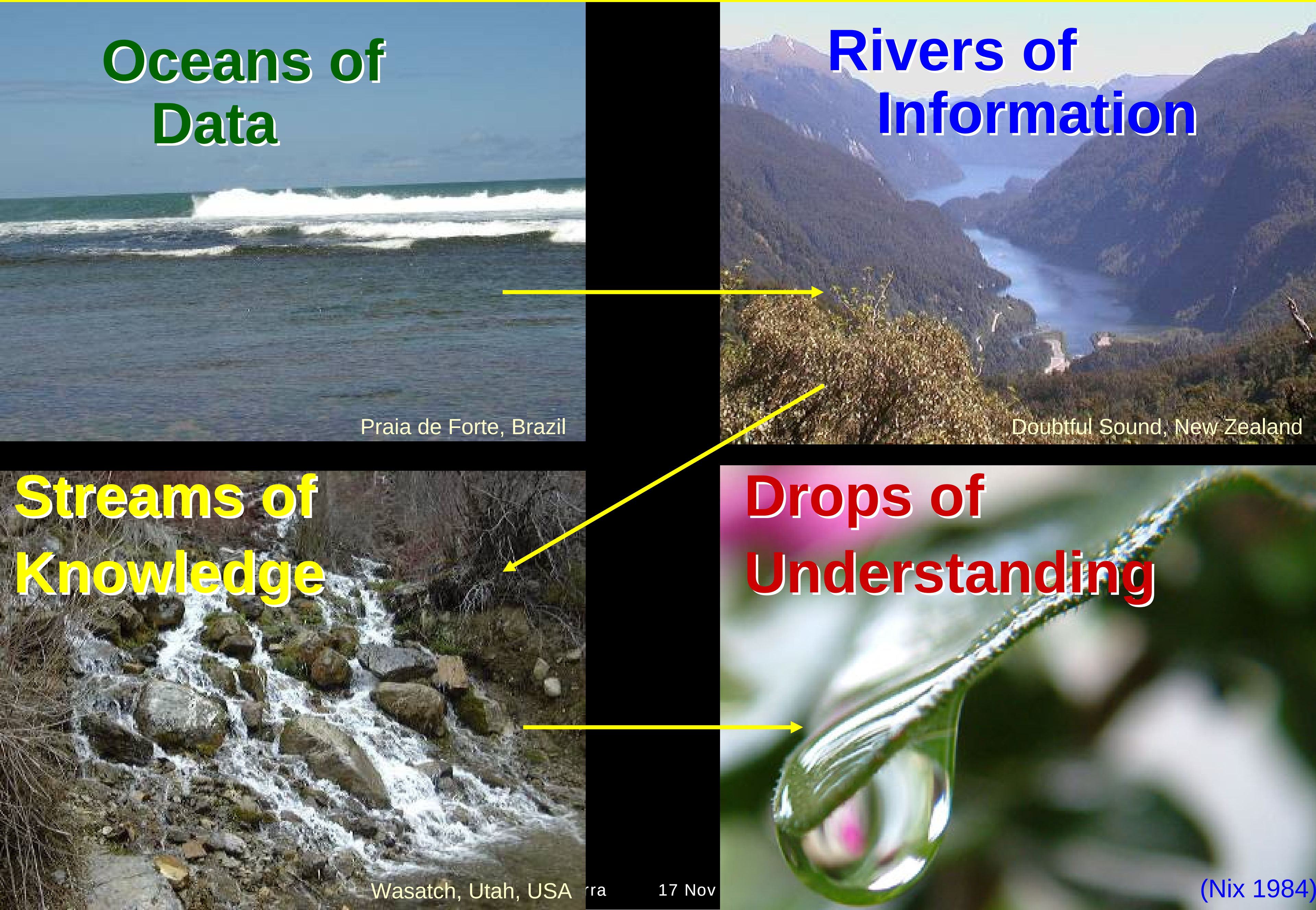
Challenges

- Many challenges – include:
  - Finding and accessing data
  - Making data useful
  - Dealing with legacy data
  - Improving Data Quality
  - Dealing with sensitive data

Great Egret, Louisiana 2004

The data equation

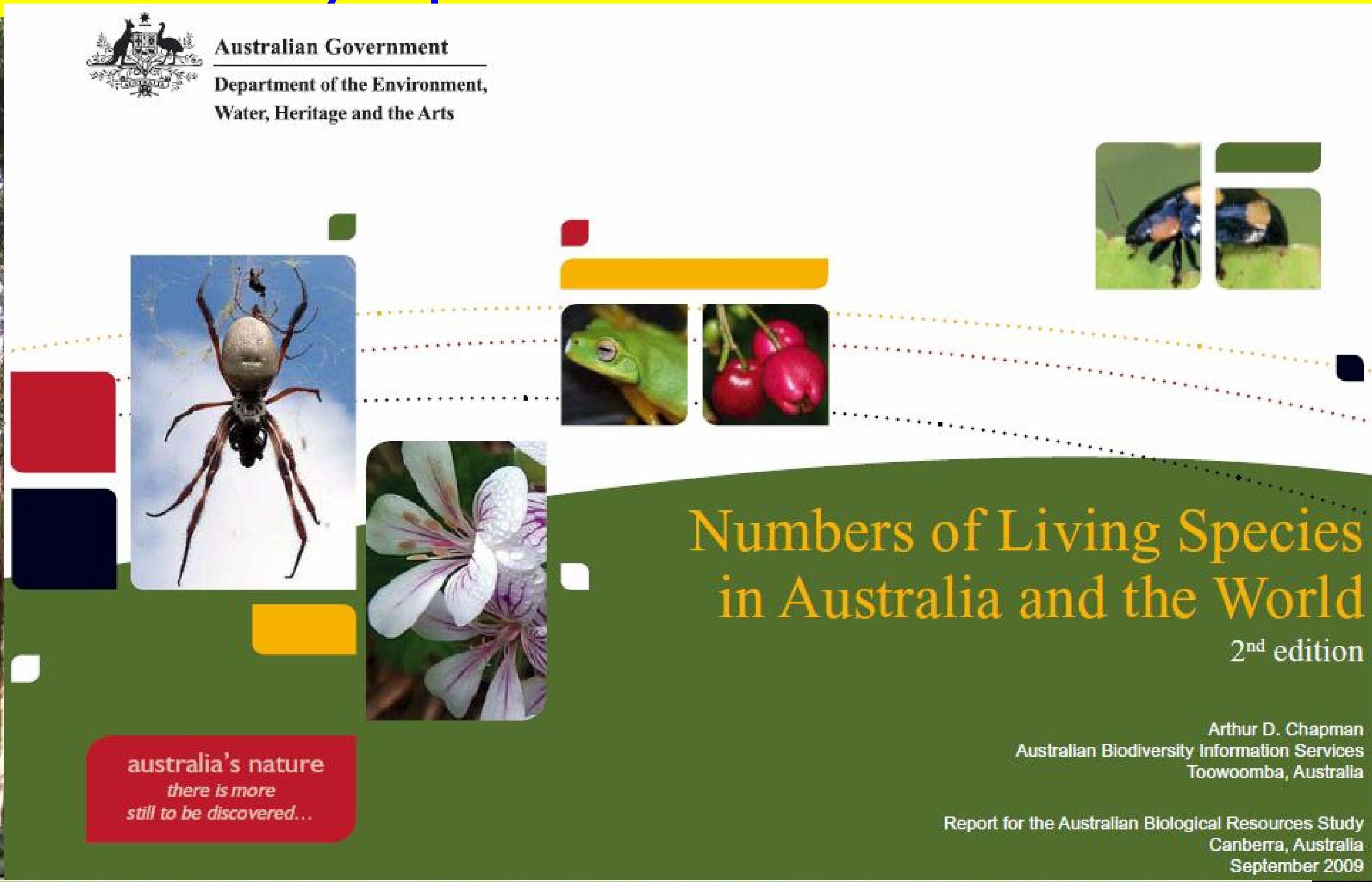


Challenges

– Finding and accessing data

Great Egret, Louisiana 2004

How many species are there?



Taxon	World Descr./ Accepted	Australia Descr./ Accepted	Austral. Percent.	Estimate World	Estimate Australia	World Threat.[14]	World Threat. Percent.	Aust Threat.[15]	Austral Threat. Percent	% of World's Threat.	Percent. Endemic
Chordates	64,788	~8,128	12.5%	~80,500	~9,088	5,966	9.2%	246	3.0%	4.1%	41.3%
Invertebrates	1,359,365	98,703	7.3%	~6,755,830	~320,465	2,524	0.2%	32	0.04%	1.3%	unknown
Plants	310,129	24,716[16]	7.9%	~390,800	26,845	8,457	2.7%	1,263	5.1%	14.9%	86%
Fungi	98,998	11,846	11.9%	1,500,000	50,000	3	>0%	0	0%	0%	unknown
Others	~66,307	>4,186	6.2%	2,600,500	~160,000	6	0.01%	0	0%	0%	unknown
TOTAL 2009	1,899,587	147,579	7.8%	~11,327,630	~566,398	16,956	0.9%	1,541	1.1%	9.1%	unknown

http://www.environment.gov.au/biodiversity/abrs/publications/other/species-numbers/2009/pubs/nlsaw-2nd-complete.pdf

Australian Biodiversity Information Services

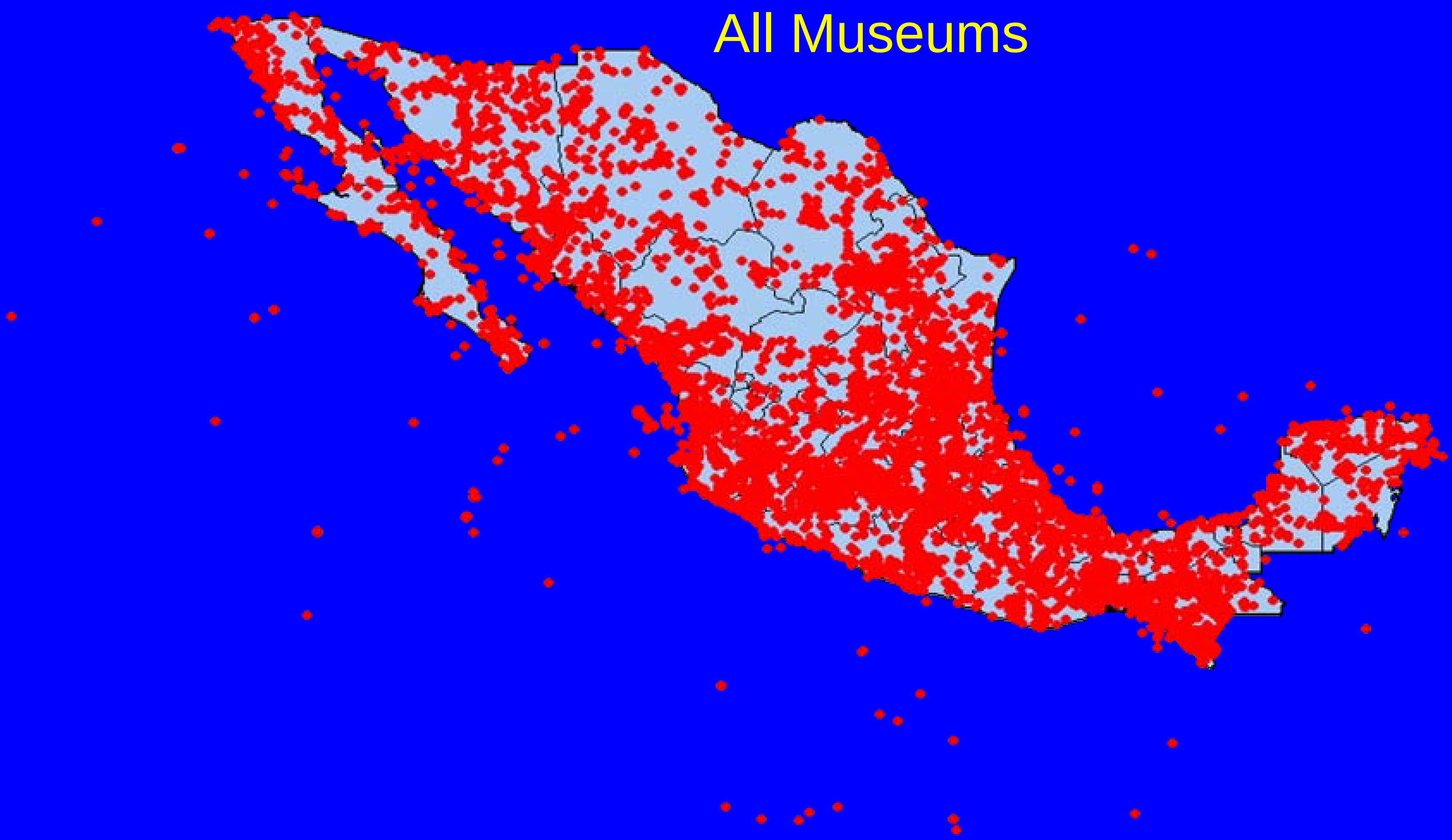
Challenges

– Making data useful


Great Egret, Louisiana 2004

Taking data to information

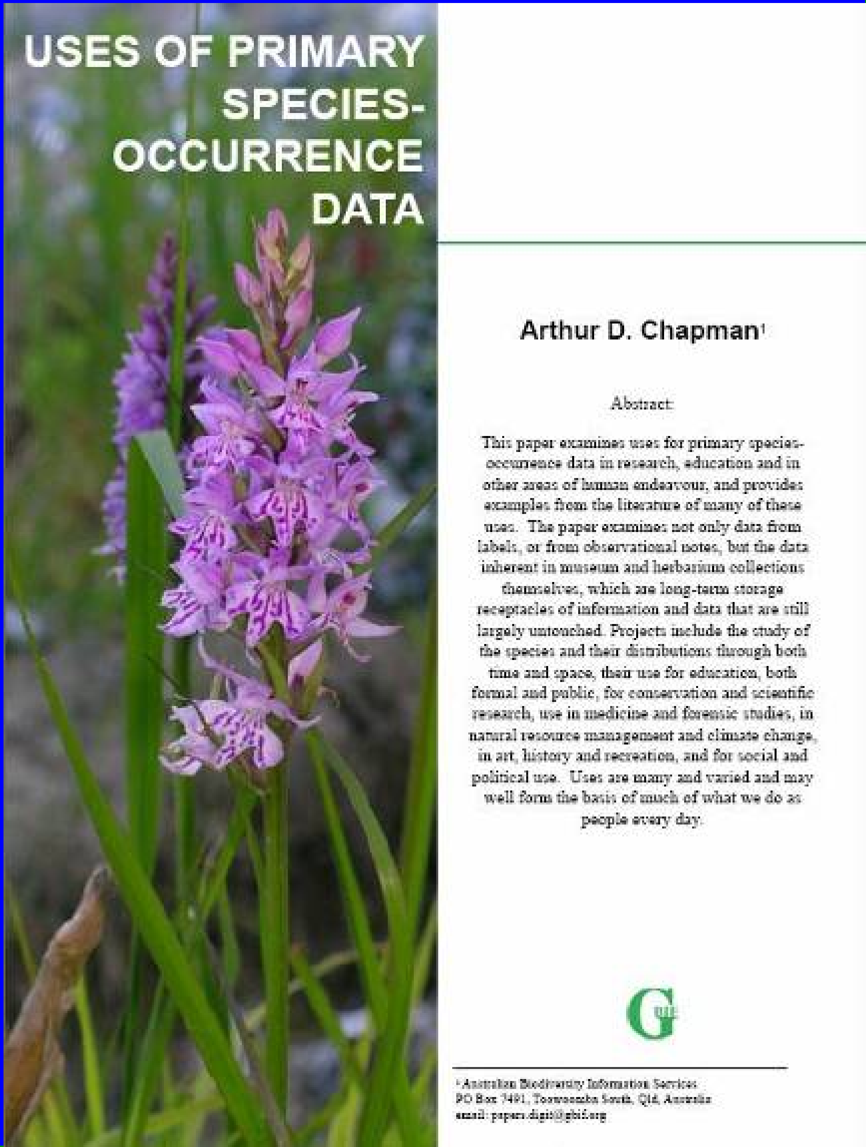
Distributed studies using mexican birds



From Beach 2001

Using species data

Taxonomic Studies, Ecological Biogeography, Phylogenies
Biogeographic Studies, Species Modelling
Species Diversity and Population studies
Life Histories and Phenologies
Studies of Threatened and Migratory species
Climate Change Impacts
Ecology, Ecosystems, Evolution and Genetics
Environmental Regionalisations
Conservation Planning
Natural Resource Management
Agriculture, Forestry, Fisheries and Mining
Health and Public Safety
Bioprospecting
Forensics
Border Control and Wildlife Trade
Education and Public Outreach
Ecotourism
Art and History, Science and Politics
Recreation
Human Infrastructure Planning



USES OF PRIMARY SPECIES-OCCURRENCE DATA

Arthur D. Chapman[1]

Abstract

This paper examines uses for primary species-occurrence data in research, education and in other areas of human endeavour, and provides examples from the literature of many of these uses. The paper examines not only data from labels, or from observational notes, but the data inherent in museum and herbarium collections themselves, which are long-term storage receptacles of information and data that are still largely untouched. Projects include the study of the species and their distributions through both time and space, their use for education, both formal and public, for conservation and scientific research, use in medicine and forensic studies, in natural resource management and climate change, in art, history and recreation, and for social and political use. Uses are many and varied and may well form the basis of much of what we do as people every day.

[1] Australian Biodiversity Information Services
PO Box 7491, Toowoomba South, Qld, Australia
email: papers.digit@gbif.org

Challenges

– Dealing with legacy data


Great Egret, Louisiana 2004

Georeferencing - The Past

- Legacy data
  - Lots of data
  - Very little data with georeferencing
  - Georeferencing added from paper maps
  - Little accuracy recording or data quality checking

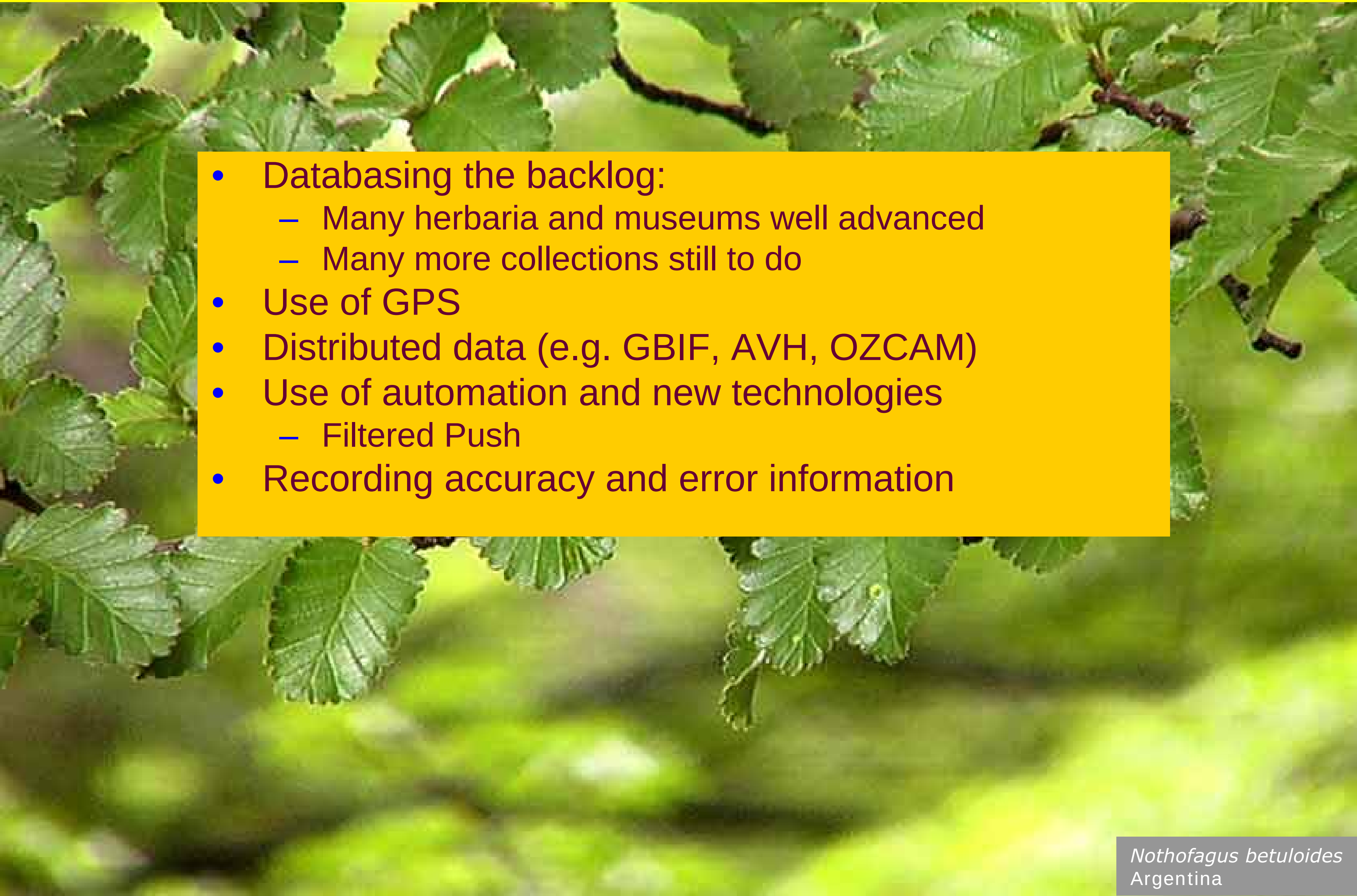
Plant and animal specimen data held in museums provide a vast information resource, providing not only present day information on the locations of these entities, but also historic information going back several hundred years

(Chapman and Busby 1994).

Fern, Brazil

Georeferencing – the present

- Databasing the backlog:
  - Many herbaria and museums well advanced
  - Many more collections still to do
- Use of GPS
- Distributed data (e.g. GBIF, AVH, OZCAM)
- Use of automation and new technologies
  - Filtered Push
- Recording accuracy and error information



Nothofagus betuloides
Argentina

Recording Accuracy and Uncertainty

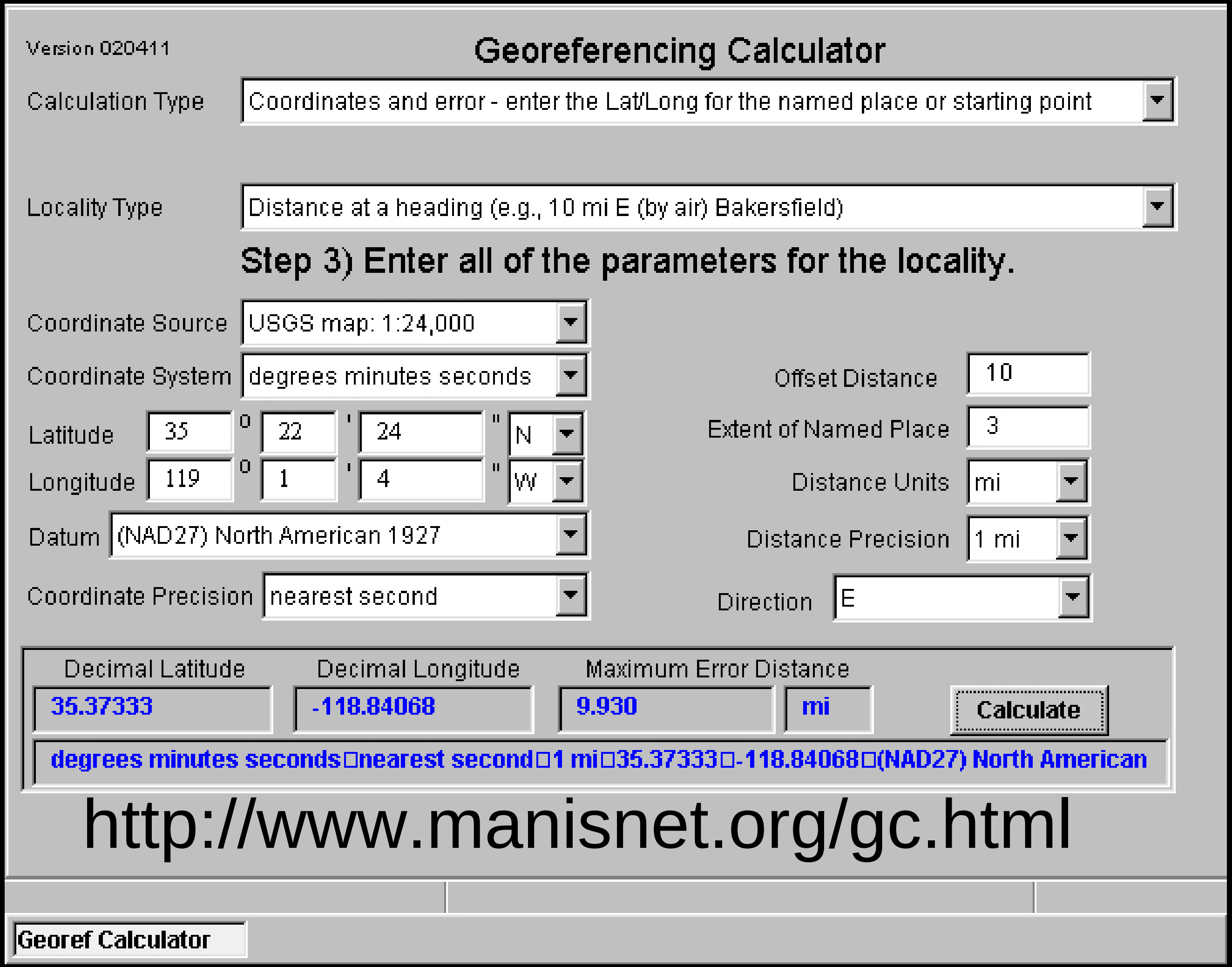
Additional Uncertainty Fields

- Preferably in meters (Point-Radius)
- Remarks

Documenting Validation tests

- Who
- What
- How

MaNiS Georeferencing Calculator



http://www.manisnet.org/gc.html

Locality Types- 1

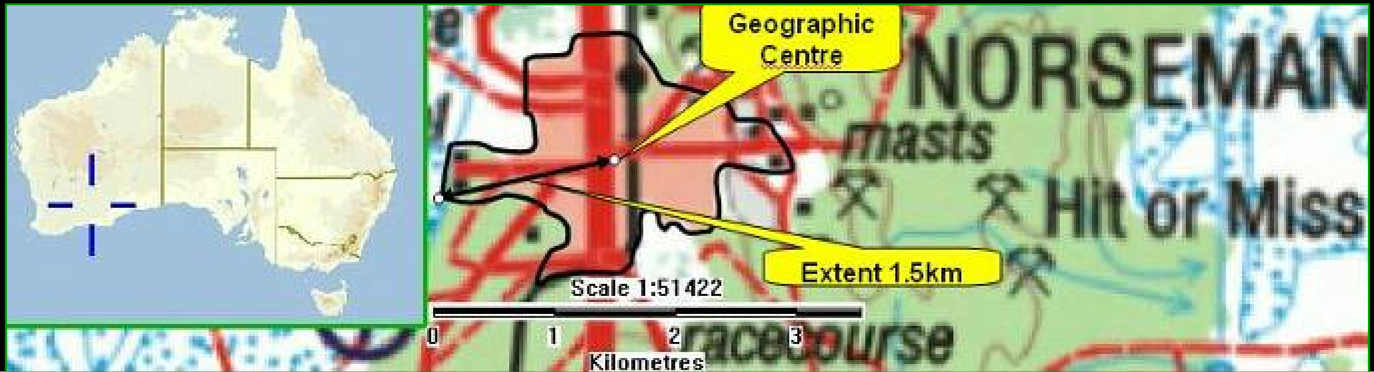
Named Place

- » **Example 1: “Point Lookout”**
- » **Example 2: “Bennetts Waterhole”**
- » **Example 3: “Isla Tiburon”**
- » **Example 4: “Lorne Reef”**
- » **Example 5: “Mt Hypipamee”**

Georeference: Use centre of named place

Extent: Use the distance from the coordinates of the named place to the furthest point within the named place

Uncertainty: Use the *MaNIS Georeference Calculator*



Australian Biodiversity Information Services

Example

Locality: "Bakersfield"
Suppose the coordinates for Bakersfield came from the GNIS database (a gazetteer) and the distance from the center of Bakersfield to the furthest city limit is 3 km.

Coordinate System: degrees minutes seconds
Latitude: 35º 22' 24" N
Longitude: 119 º 1' 4" W
Datum: not recorded; 79 m uncertainty
Coordinate Precision: nearest second; 40 m uncertainty
Coordinate Source: gazetteer
Extent of Named Place: 3 km
Distance Units: km

Decimal Latitude: 35.37333
Decimal Longitude: -119.01778
Maximum Uncertainty Distance: 3.119 km

Australian Biodiversity Information Services

Locality Types - 2

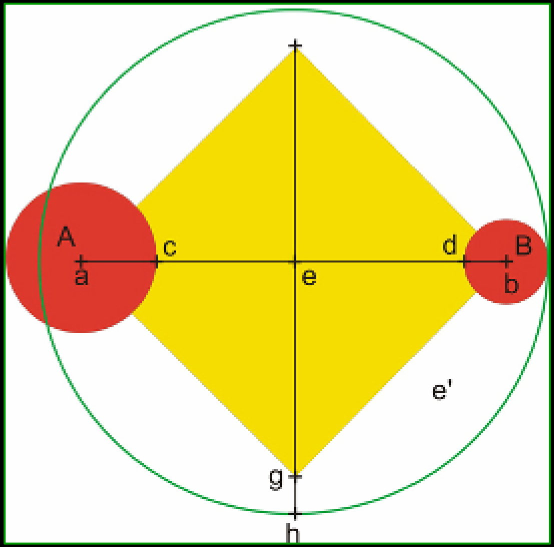
Between two Named Places

» **Example 1:** "between Point Reyes and Inverness"

Georeference: Find the coordinates of the midpoint between the centres of the two named places **(e)**

Extent: Use the extent of A or B, whichever is greater, plus one-half the distance between the centres of A and B.

Uncertainty: Use the ***MaNIS Georeference Calculator***
Calculate the same as for '***Named Place***'.



Australian Biodiversity Information Services

BioGeomancer

APPLICATIONS STANDARDS LIBRARY NEWS

Home About Feedback

WHAT IS THE BIOGEOMANCER PROJECT?

The BioGeomancer (BG) Project is a worldwide collaboration of natural history and geospatial data experts. The primary goal of the project is to maximize the quality and quantity of biodiversity data that can be mapped in support of scientific research, planning, conservation, and management. The project promotes discussion, manages geospatial data and data standards, and develops software tools in support of this mission. **Learn more about us**!

WHAT IS GEOREFERENCING?

Georeferencing is the process of converting text descriptions of locations to computer-readable geographic locations, such as a GIS system uses. **More about georeferencing**...

LATEST BIOGEOMANCER NEWS

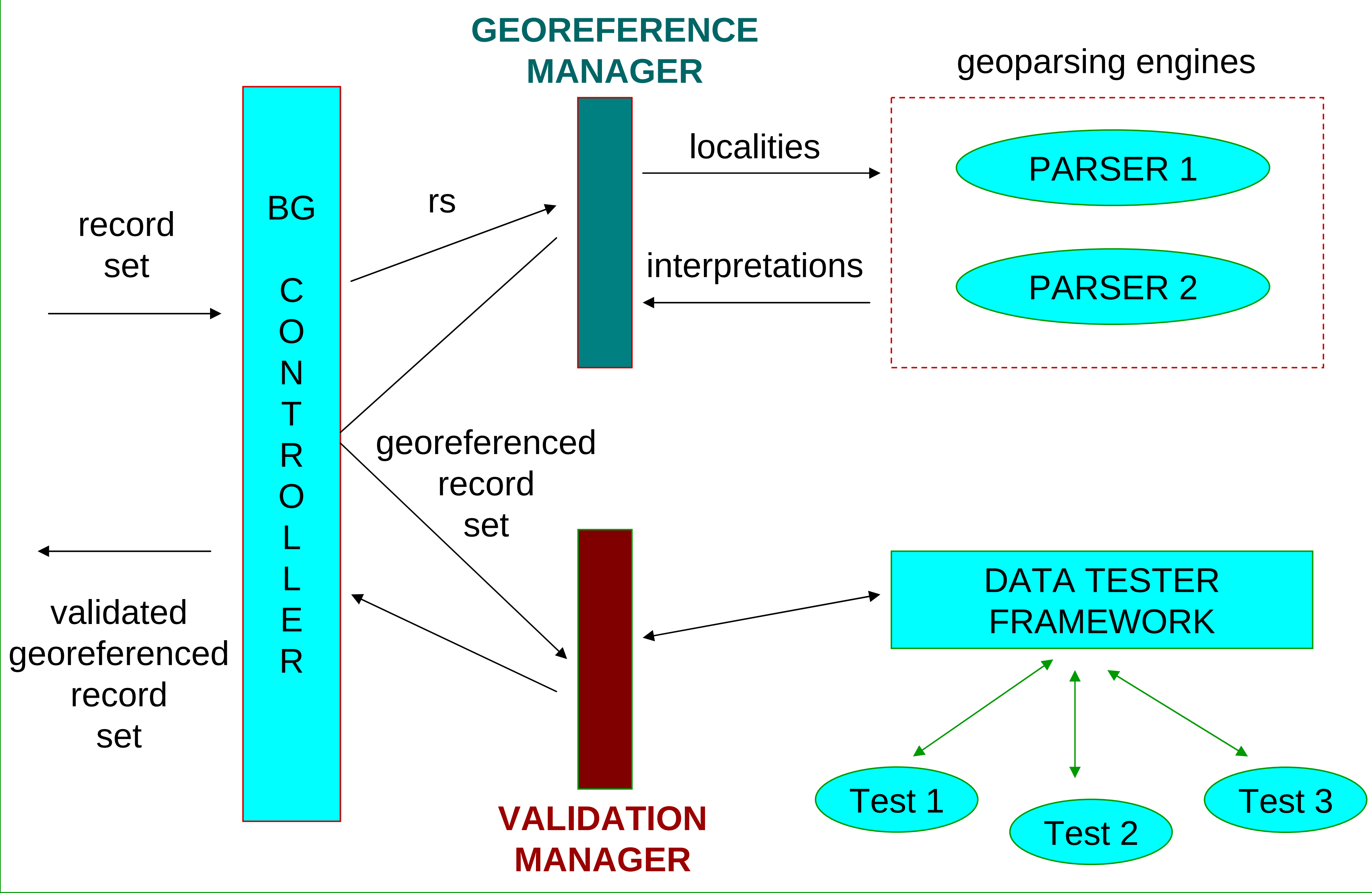
BioGeomancer features in August edition of GEOWORLD!

More News...

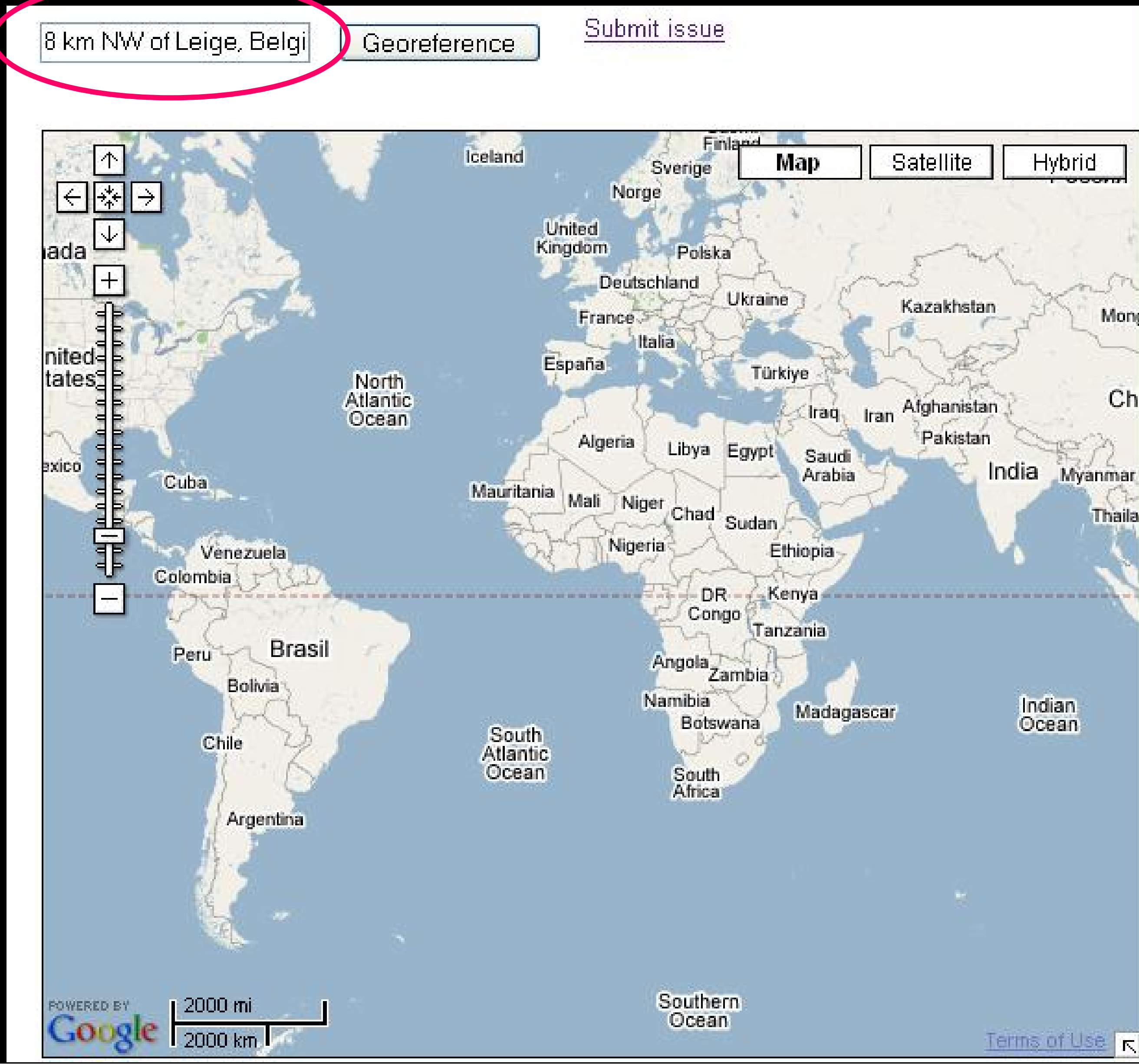
www.biogeomancer.org

© Moore Foundation 2005

BioGeomancer – Georeferencing



BioGeomancer Workbench



BioGeomancer

8 km NW Liege, Belgium | Georeference | Submit issue

B

Belgium: 8 km NW of Liège

50.6841853, 5.4867103

2447 meter uncertainty

BioGeomancer Workbench – Batch Georeferencing

- Need to register and Log in
- Submit a project
  - Use the “Help” to see how to go about submitting files
  - (See next slide for formatting)
  - NB – files must be in UTF-8
- Results downloaded as XML file

Records	Project ID	Project name			
32 records	5629	Training-NL	Download	Create Features Project	Delete

Australian Biodiversity Information Services

Format for batch georeferencing

How do I format a file for batch georeferencing?

Save your data in a tab-delimited text file.The first row of the must contain column names. BioGeomance
can understands concept names from the DarwinCore 1.4 and its Geospatial Extension.

Specifically, BioGeomancer currently interprets the following fields (the order and case are not relevant):

Locality
HigherGeography
Continent
WaterBody
IslandGroup
Island
Country
StateProvince
County
VerbatimLatitude
VerbatimLongitude
VerbatimCoordinates
VerbatimElevation

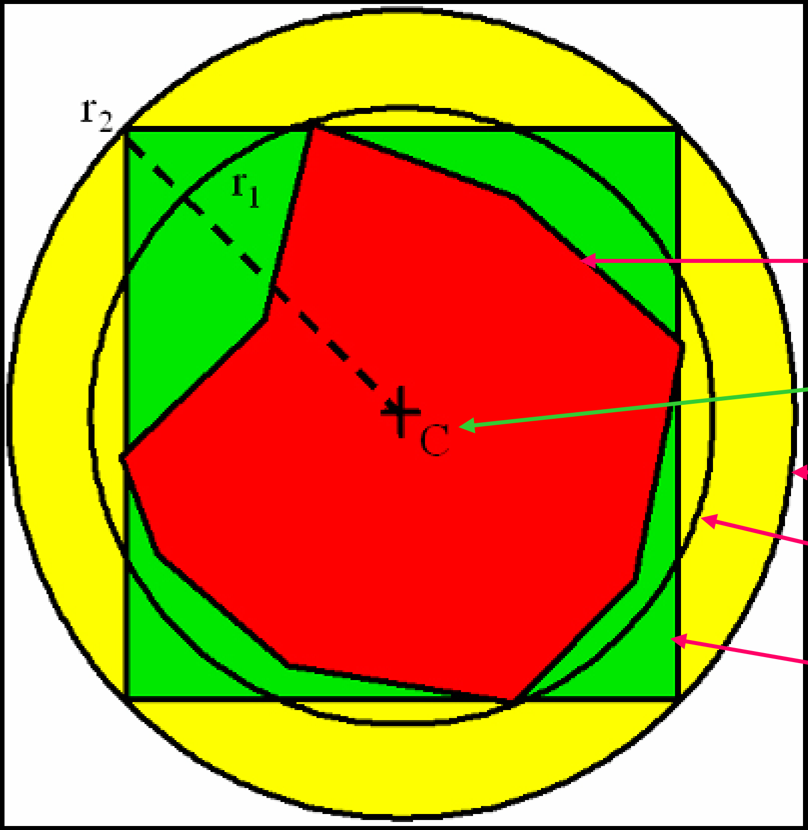
Any other fields submitted in the upload are retained, but uninterpreted. Of those in the list above, the Verbatim fields are currently unused in the spatial description, and the WaterBody, IslandGroup, and Island are likely to be heavily under-represented in the gazetteer.

Pay attention to the encoding system used. You should use UTF-8, a character encoding system for the Unicode standard to represent any of the world's scripts. Text encoded in the standard windows/English "Latin-1" will not be corrupted if it contains pure ASCII characters.

ustralian
odiversity
formation
rvices

Spatial Fit

A measure of how well the geometric representation matches the original spatial representation.



For an area where the original spatial representation of a locality is the **red polygon** with area ‘A’. The spatial fit is:

1.0

0

$(Pi*r_2^2)/A$

$(Pi*r_1^2)/A$

$(2*r_2^2)/A$

From J. Wieczorek, in Chapman and Wieczorek (eds) (2006)

Challenges

– Improving Data Quality

Great Egret, Louisiana 2004

Users need quality information

So what do we mean by 'Data Quality'?

An essential or distinguishing characteristic necessary for [spatial] data to be fit for use.

SDTS 02/92

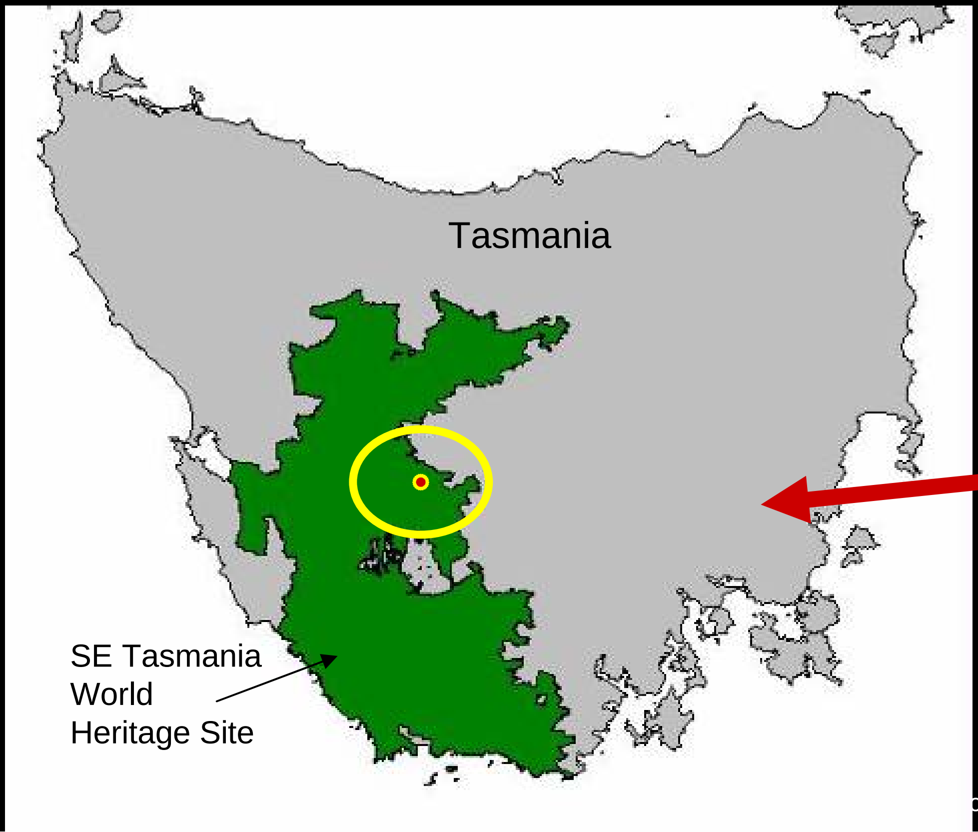
The general intent of describing the quality of a particular dataset or record is to describe the fitness of that dataset or record for a particular use that one may have in mind for the data. (Chrisman 1991)

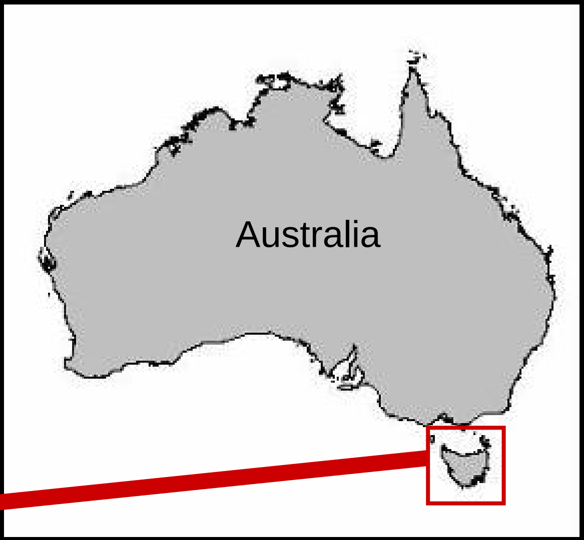
Data quality - fitness for use?

Fitness for use

Does species 'A' occur in Tasmania?

Does species 'A' occur in National Park 'y'





Australian Biodiversity Information Services

Loss of data quality

Loss of data quality can occur at many stages:

- At the time of collection
- During digitization
- During documentation
- During storage and archiving
- During analysis and manipulation
- At time of presentation
- And through the use to which they are put

Don’t underestimate the simple elegance of quality improvement. Other than teamwork, training, and discipline, it requires no special skills. Anyone who wants to can be an effective contributor.

(Redman 2001).

Australian Biodiversity Information Services

Principles of data quality

It is important for organizations to have

a vision with respect to having good quality data.

a policy to implement that vision, and

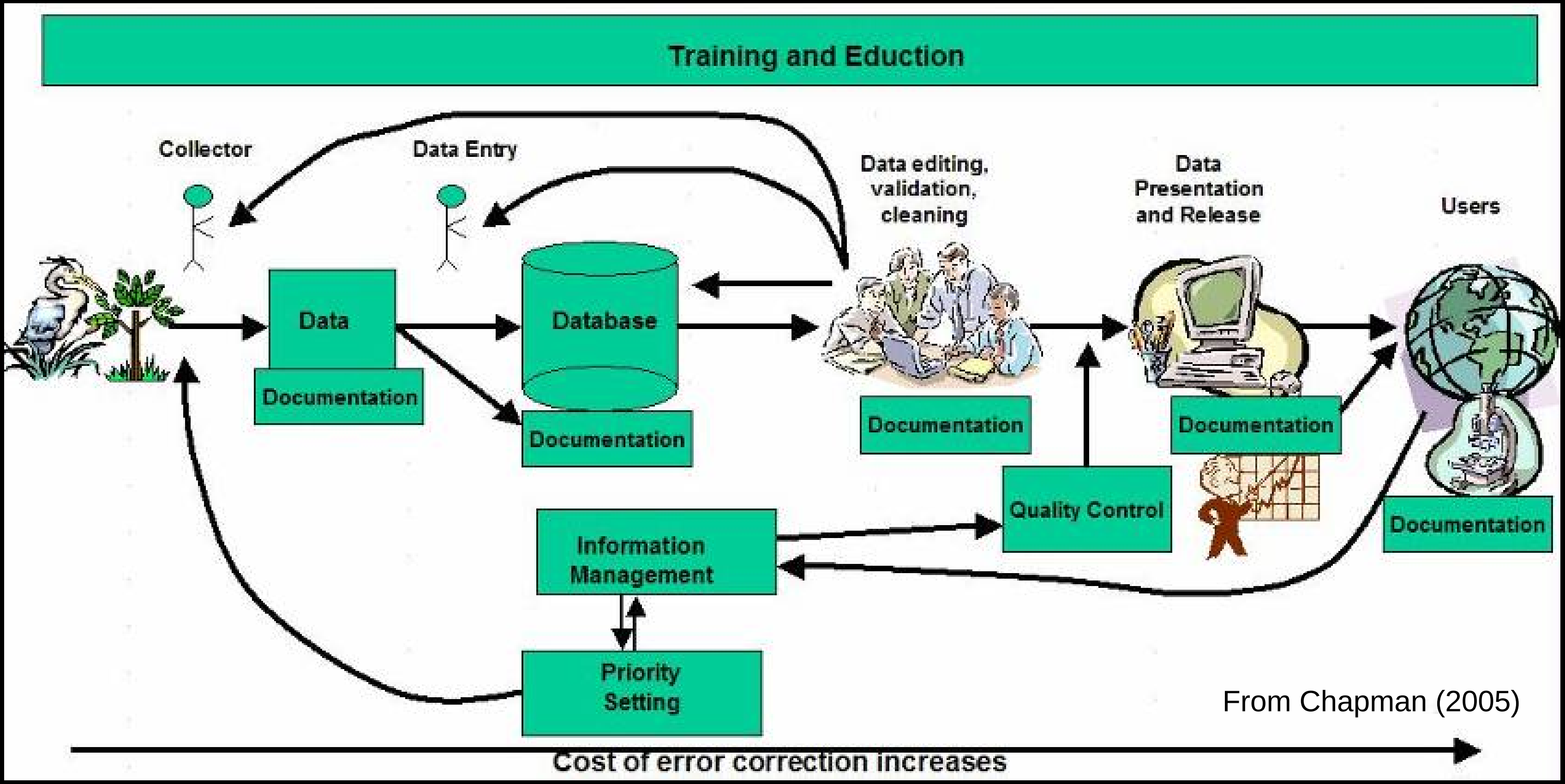
a strategy for implementation

Experience has shown that treating data as a long-term asset and managing it within a coordinated framework produces considerable savings and ongoing value.

(NLWRA 2003).

Australian Biodiversity Information Services

Data Quality Information Chain



Assign responsibility for the quality of data to those who create them. If this is not possible, assign responsibility as close to data creation as possible

(Redman 2001)

Australian Biodiversity Information Services

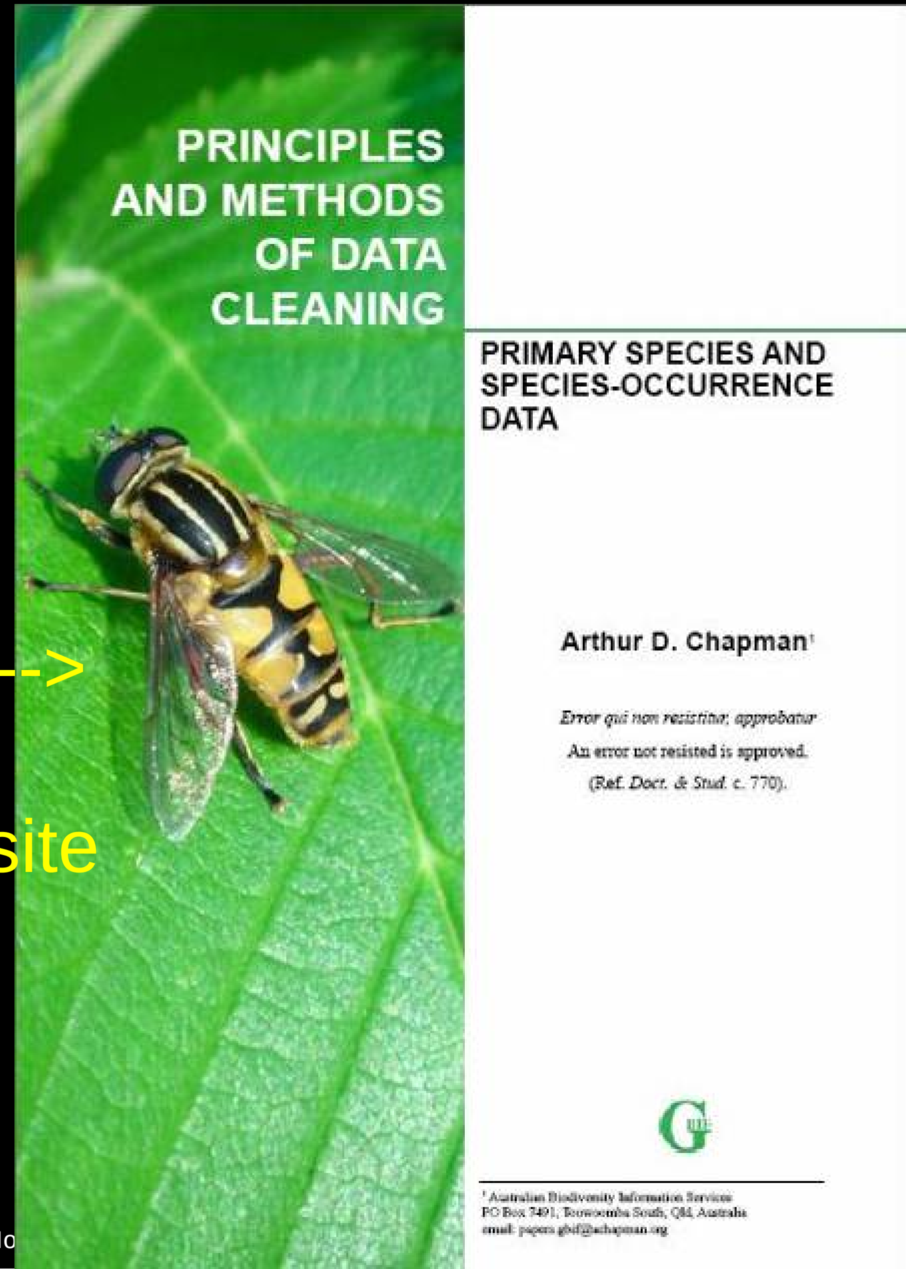
Data validation

Two key sources of error are:

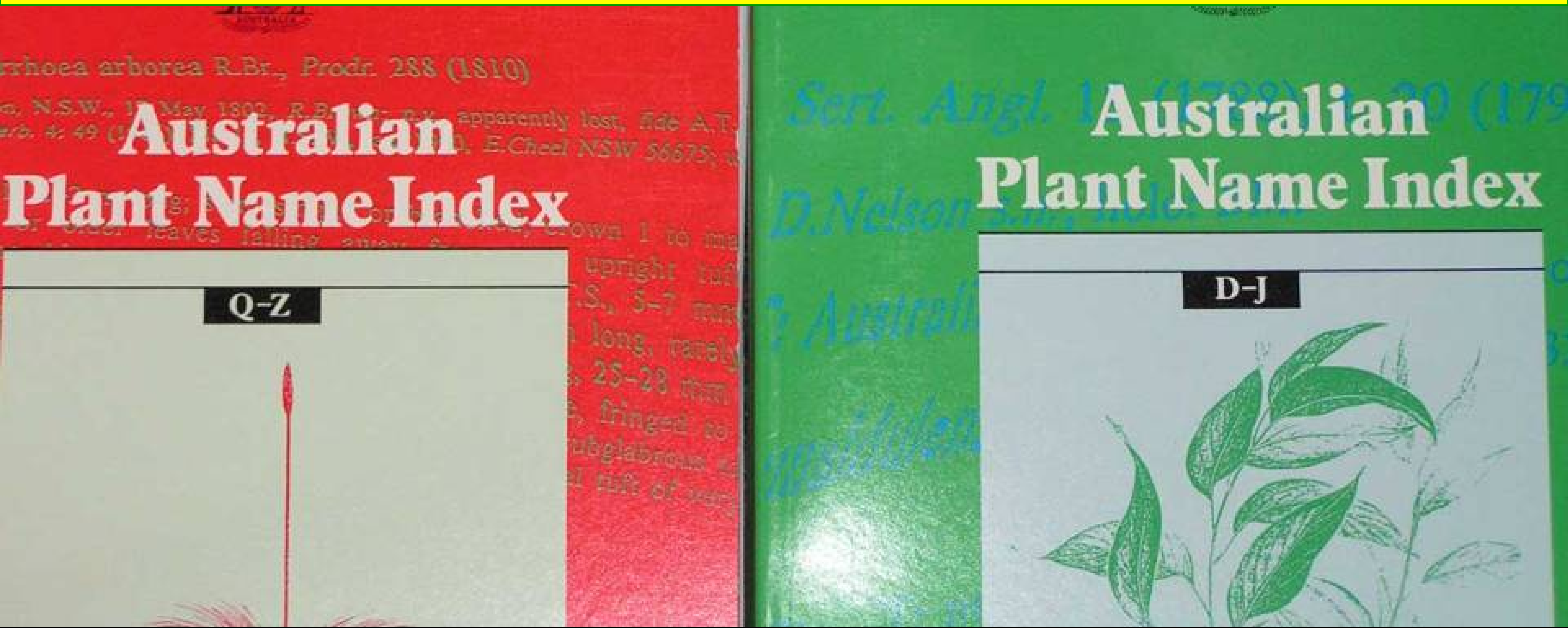
- Taxonomic names
- Georeferences (lat's and long's)

Methods for identifying error Documented here -------------->

available via GBIF web site
http://www.gbif.org



1. Taxonomic and Nomenclature Data



Chapman (1991) http://www.cpbr.gov.au/cpbr/databases/apni.html

Taxonomic Data

Consists of: (not all are always present):

- Name (scientific, common, hierarchy, rank)
- Nomenclatural status (synonym, accepted, typification)
- Reference (author, place and date of publication)
- Determination (by whom and when the record was identified)
- Type specimen citation
- Quality fields (accuracy of determination, qualifiers)

Documenting Taxonomic Data Quality

- Several methods exist for documenting taxonomic verification - none are completely satisfactory
  - Herbarium Information Standards and Protocols for the Interchange of Data (HISPID)
  - Australian National Fish Collection (1993)
  - Several others restricted to one or two institutions

- Proposal – four level:
  - Who determined the specimen and when
  - What was used (type specimen, local flora, monograph, etc.)
  - Level of expertise of the determiner
  - What confidence did the determiner have in the determination.

Taxon Verification Status - proposed

Name of determinor:

Date of determination:

Source of determination: (e.g. compared with holotype, used national flora)

- identified by **World expert** in the taxon with **high certainty**
- identified by **World expert** in the taxon with **reasonable certainty**
- identified by **World expert** in the taxon with **some doubt**
- identified by **regional expert** in the taxon with **high certainty**
- identified by **regional expert** in the taxon with **reasonable certainty**
- identified by **regional expert** in the taxon with **some doubt**
- identified by **non-expert** in the taxa taxon **high certainty**
- identified by **non-expert** in the taxa taxon **reasonable certainty**
- identified by **non-expert** in the taxa taxon **some doubt**
- identified by **the collector** with **high certainty**
- identified by **the collector** with **reasonable certainty**
- identified by **the collector** with some doubt.

From: Chapman (2005) Principles of Data Quality. GBIF

Australian Biodiversity Information Services

Error checking

- **Missing Data Values**
  - empty fields where values should occur

    (e.g. if a species epithet is present, then a generic name MUST be present)

Australian Biodiversity Information Services

Error checking

- **Incorrect Data Values**
  - typographic errors,
  - transposition of key strokes,
  - data entered in the wrong place

    (e.g. a species epithet present in a generic name field)

Can often be identified using Soundex/Phonex techniques

Error checking

- **Nonatomic Data Values**
  - More than one fact entered into a single field

    (e.g. a species bionomial or trinomial present in a single field)

Australian Biodiversity Information Services

Error checking

- **Domain Schizophrenia**
  - – Fields used for purposes for which they weren't intended

Family	Genus	Species
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	To be determined

e.g.

Good reference:

Dalcin, E.C. 2004. *Data Quality Concepts and Techniques Applied to Taxonomic Databases.*

Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. November 2004. 266 pp.

http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf

CRIA Data Cleaning

http://splink.cria.org.br/dc

speciesLink

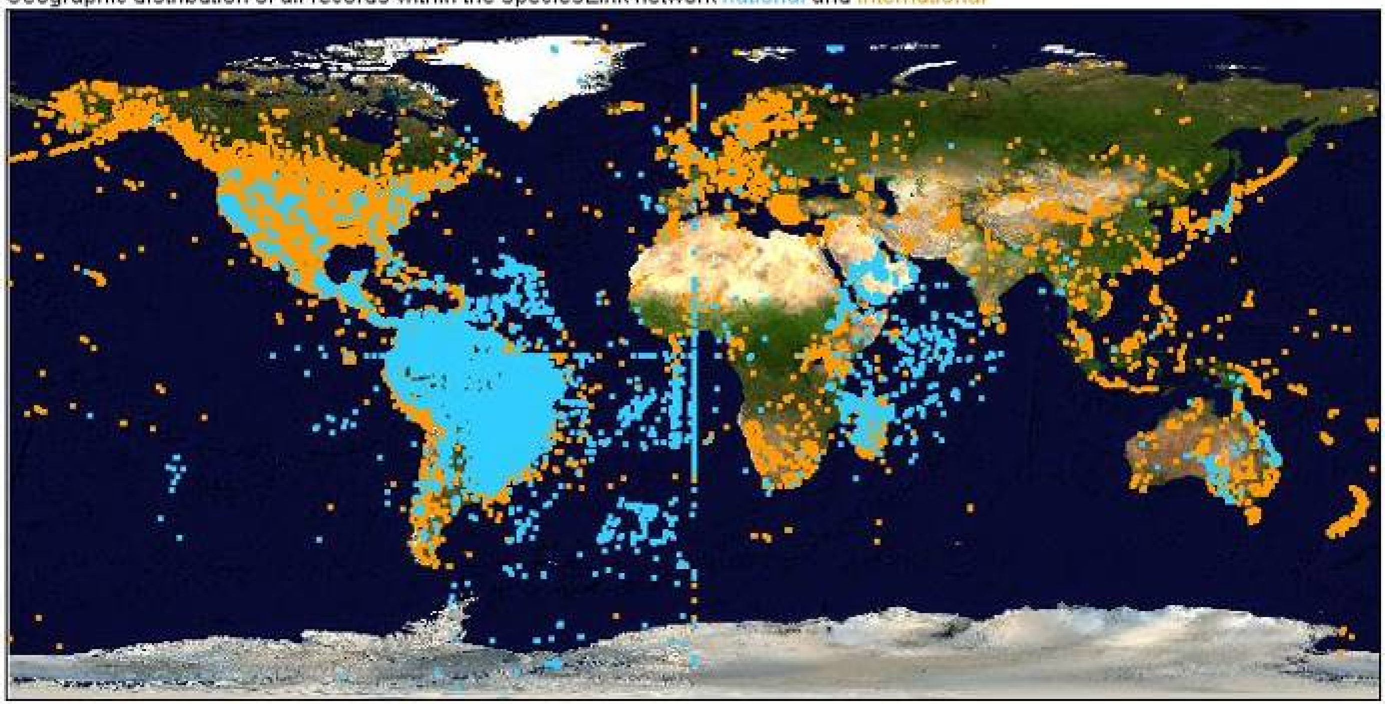
data & tools

data cleaning

This tool aims at helping curators in identifying possible errors and to standardize data. Records are not modified. The system just presents "suspect" records, recommending that they be checked by each author or curator. The tool is under constant development, so any suggestion is more then welcome.

Select a collection HSJRP

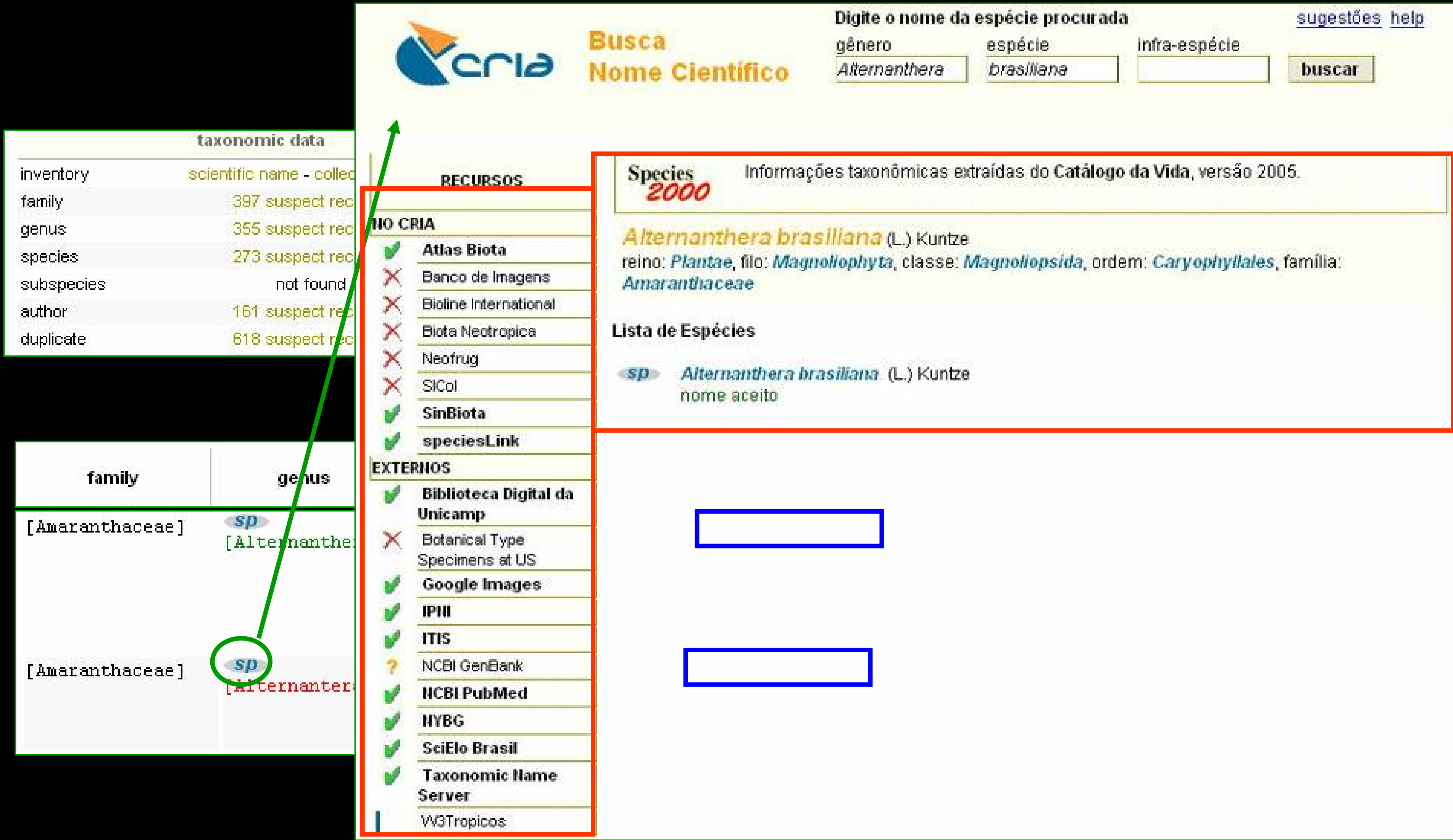
Geographic distribution of all records within the speciesLink network national and international



graphic representation of families
graphic representation of Brazilian states
origin of the records
main collectors
collection events by year

general graphic of all national collections
general graphic of all international collections

CRIA Data Cleaning



CRIA Data Cleaning

taxonomic data	
inventory	scientific name - collector - types
family	397 suspect records
genus	355 suspect records
species	273 suspect records
subspecies	not found
author	161 suspect records
duplicate	618 suspect records

sp	[Inga]	[cylindrica]	[]	2	see	51	accepted name	32009 64192
sp	[Inga]	[cilindrica]	[]	1	see	2		50335
sp	[Inga]	[cilyndrica]	[]	0		1		
sp	[Ipomoea]	[regnellii]	[]	2	see	7		29867 31014
sp	[Ipomoea]	[regnelli]	[]	1	see	1		115815

CRIA Data Cleaning

taxonomic data	
inventory	scientific name - collector - types
family	907 suspect records
genus	1074 suspect records
specie	935 suspect records
subspecie	not found
author	3060 suspect records
duplicate	197 suspect records

genus	species	subspecies	author	ocor_col	ocor_total
[Acacia]	[martusiana]	[]	[Burk.]	3	3
[Acacia]	[martusiana]	[]	[(Steud.) Burk]	1	1
[Acacia]	[martusiana]	[]	[(Steud.) Burk.]	0	2
[Actinostemon]	[concolor]	[]	[Müll.Arg.]	0	1
[Actinostemon]	[concolor]	[]	[(Spreng.) Muell.Arg.]	0	22
[Actinostemon]	[concolor]	[]	[(Spreng) Muell Arg]	45	45
[Actinostemon]	[concolor]	[]	[(Spreng.) Mull. Arg.]	1	2
[Actinostemon]	[concolor]	[]	[(spreng.) Müll. Arg.]	0	2
[Actinostemon]	[concolor]	[]	[(Spreng.) Müll. Arg.]	0	23
[Actinostemon]	[concolor]	[]	[(Spreng.) Müll.Arg.]	0	65
[Actinostemon]	[concolor]	[]	[(Spreng) Müll. Arg.]	0	6
[Actinostemon]	[concolor]	[]	[(spr.) Muell. Arg.]	0	1
[Actinostemon]	[concolor]	[]	[(Spr.) Muell.Arg.]	0	2

Australian Biodiversity Information Services

CRIA Data Cleaning

taxonomic data	
inventory	scientific name - collector - types
family	907 suspect records
genus	1074 suspect records
specie	935 suspect records
subspecie	not found
author	3060 suspect records
duplicate	197 suspect records

collector	collector number	collectioncode	genus	species	subspecies	identified by	ocor_col
Aguiar, O.T.	193	ESA	sp Sapium	glandulatum		Cordeiro, I.	1
Aguiar, O. T.	193	UEC	sp Sapium	longifolium		J. A Pastore	1
Aguiar, O.T.	193	SP	sp Sapium	glandulatum		Cordeiro, I.	1
Amaral Jr, A	118	UEC	sp Miconia	cubatanensis			1
Amaral Jr, A	118	UEC	sp Miconia	cubatanensis		Goldenberg, R	1
Amaral Jr, A.	118	SPF	sp Picramnia	sellowii	subsp. sellowii		1
Amaral Jr., A.	118	SPF	sp Picramnia	sellowii	subsp. sellowii		1

Australian Biodiversity Information Services

CRIA data cleaning

speciesLink

data & tools

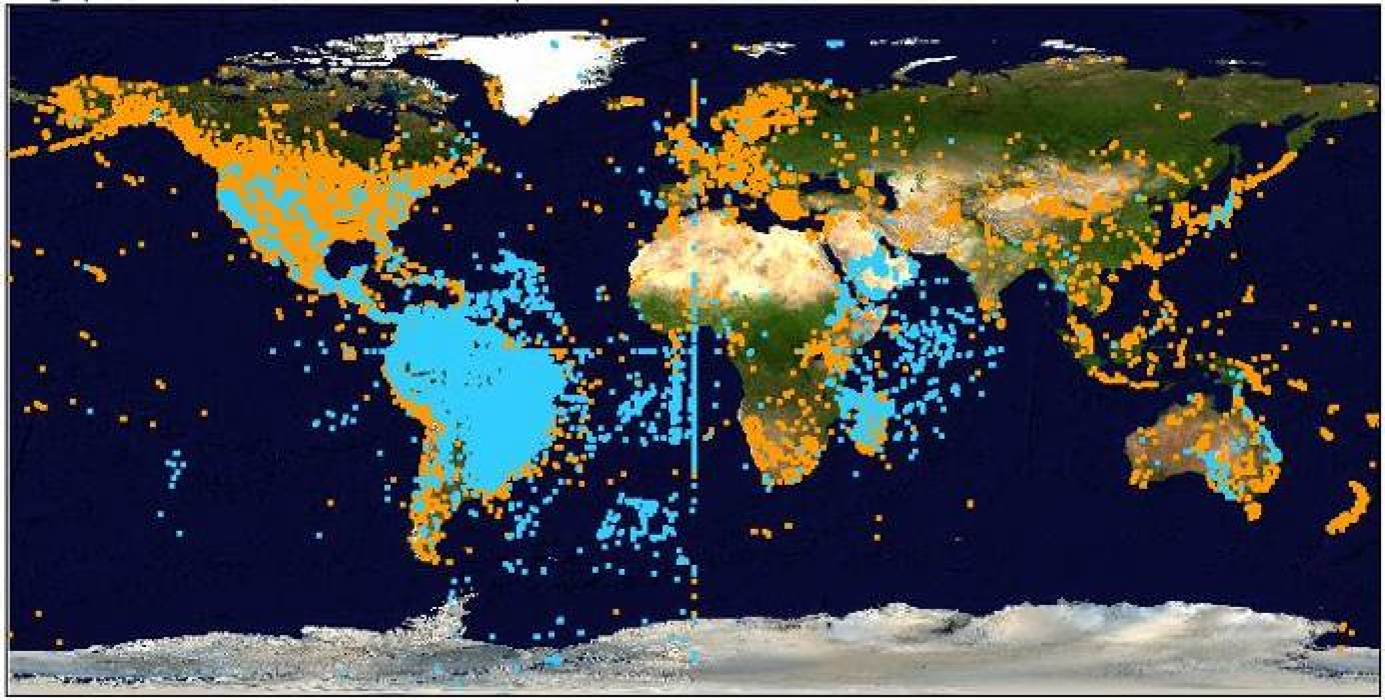
http://splink.cria.org.br/dc

data cleaning

This tool aims at helping curators in identifying possible errors and to standardize data. Records are not modified. The system just presents "suspect" records, recommending that they be checked by each author or curator. The tool is under constant development, so any suggestion is more then welcome.

Select a collection UEC

Geographic distribution of all records within the speciesLink network national and international



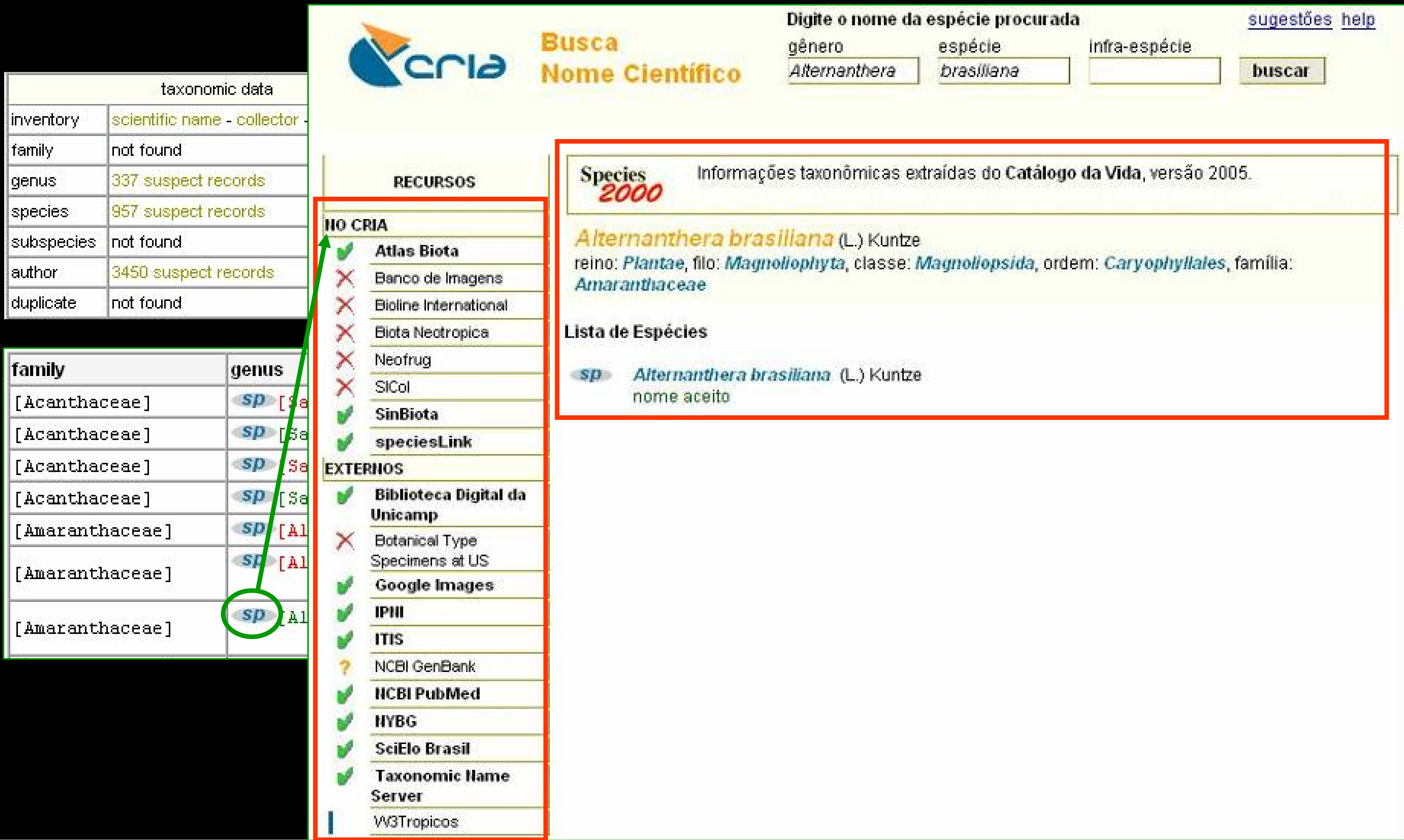
graphic representation of families
graphic representation of Brazilian states
origin of the records
main collectors
collection events by year

general graphic of all national collections
general graphic of all international collections

Australian Biodiversity Information Services

The Brazil example

http://splink.cria.org.br/dc



The Brazil example

http://splink.cria.org.br/dc

taxonomic data	
inventory	scientific name - collector - types
family	not found
genus	337 suspect records
species	957 suspect records
subspecies	not found
author	3450 suspect records
duplicate	not found

genus	species	subspecies	ocor_col	ocor_total	status_sp2000
SP [Aechmea]	[distichanta]	[]	15	16	
SP [Aechmea]	[distichantha]	[]	1	11	
SP [Aeschynomene]	[brasiliana]	[var. brasiliana]	13	13	
SP [Aeschynomene]	[brasiliiana]	[var. brasiliana]	1	1	
SP [Acalypha]	[gracilis]	[]	8	48	
SP [Acalypha]	[gracillis]	[]	8	8	
SP [Acacia]	[polyphylla]	[]	82	292	accepted name
SP [Acacia]	[poliphylla]	[]	1	1	
SP [Acacia]	[polyphilla]	[]	0	1	
SP [Acacia]	[polyphyllla]	[]	1	1	

The Brazil example

http://splink.cria.org.br/dc

taxonomic data	
inventory	scientific name - collector - types
family	not found
genus	337 suspect records
species	957 suspect records
subspecies	not found
author	3450 suspect records
duplicate	not found

genus	species	subspecies	author	ocor_col	ocor_total
sp [Acacia]	[martusiana]	[]	[Burk.]	3	3
sp [Acacia]	[martusiana]	[]	[(Steud.) Burk]	1	1
sp [Acacia]	[martusiana]	[]	[(Steud.) Burk.]	0	2
sp [Actinostemon]	[concolor]	[]	[Müll.Arg.]	0	1
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Muell.Arg.]	0	22
sp [Actinostemon]	[concolor]	[]	[(Spreng) Muell Arg]	45	45
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Mull. Arg.]	1	2
sp [Actinostemon]	[concolor]	[]	[(spreng.) Müll. Arg.]	0	2
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Müll. Arg.]	0	23
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Müll.Arg.]	0	65
sp [Actinostemon]	[concolor]	[]	[(Spreng) Müll. Arg.]	0	6
sp [Actinostemon]	[concolor]	[]	[(spr.) Muell. Arg.]	0	1
sp [Actinostemon]	[concolor]	[]	[(Spr.) Muell.Arg.]	0	2

The Brazil example

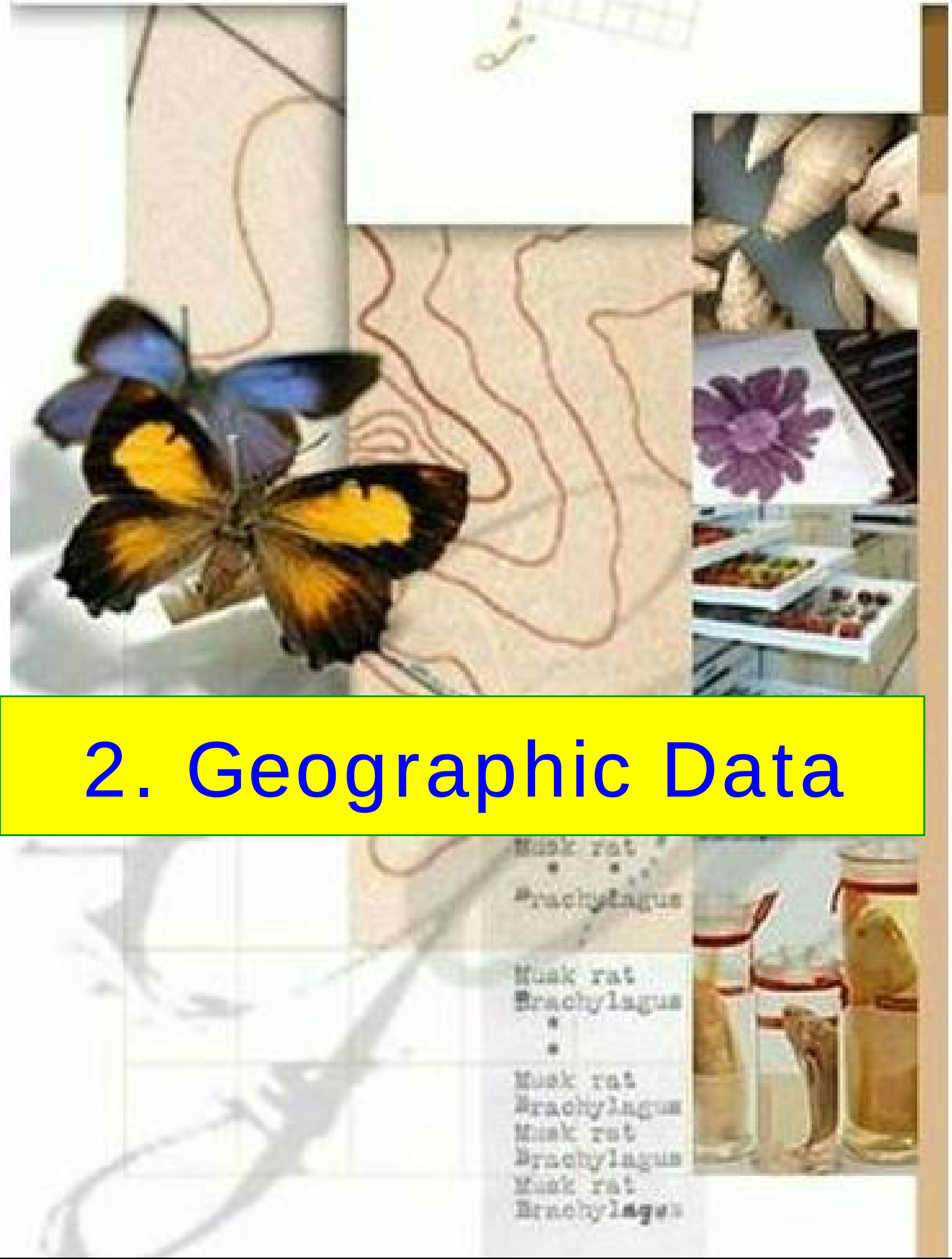
http://splink.cria.org.br/dc

taxonomic data	
inventory	scientific name - collector - types
family	907 suspect records
genus	1074 suspect records
specie	935 suspect records
subspecie	not found
author	3060 suspect records
duplicate	197 suspect records

collector	collector number	collectioncode	genus	species	subspecies	identified by	ocor_col
Aguiar, O.T.	193	ESA	sp Sapium	glandulatum		Cordeiro, I.	1
Aguiar, O. T.	193	UEC	sp Sapium	longifolium		J. A Pastore	1
Aguiar, O.T.	193	SP	sp Sapium	glandulatum		Cordeiro, I.	1
Amaral Jr, A	118	UEC	sp Miconia	cubatanensis			1
Amaral Jr, A	118	UEC	sp Miconia	cubatanensis		Goldenberg, R	1
Amaral Jr, A.	118	SPF	sp Picramnia	sellowii	subsp. sellowii		1
Amaral Jr., A.	118	SPF	sp Picramnia	sellowii	subsp. sellowii		1

Australian Biodiversity Information Services

2. Geographic Data

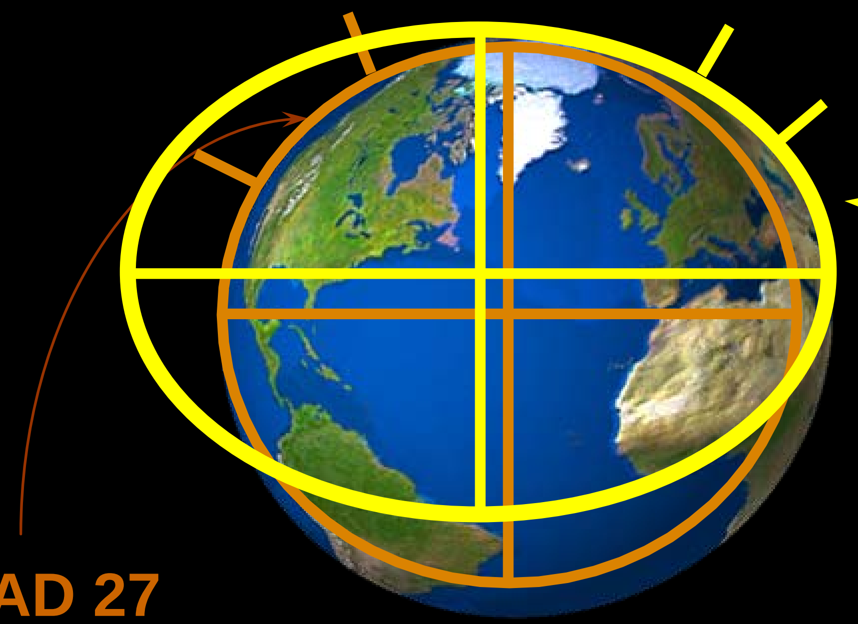


BioGeomancer

Guide to Best Practices for Georeferencing

Traditional Datums

Traditional Horizontal Datums



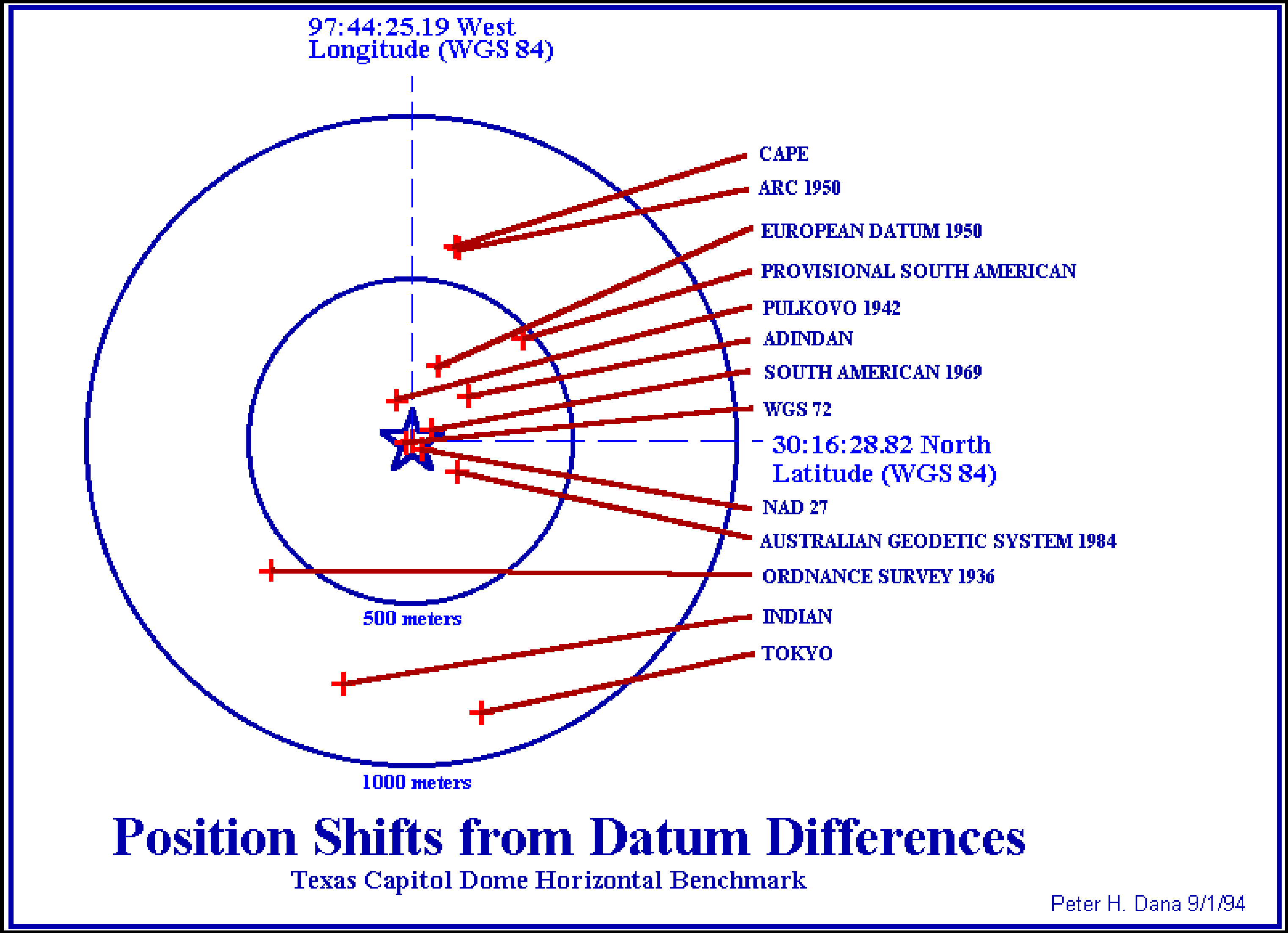
NAD 27
(Clarke Ellipsoid)

ED 50
(International Ellipsoid)

From US Navy (n.dat.)

Australian Biodiversity Information Services

Datum Shifts



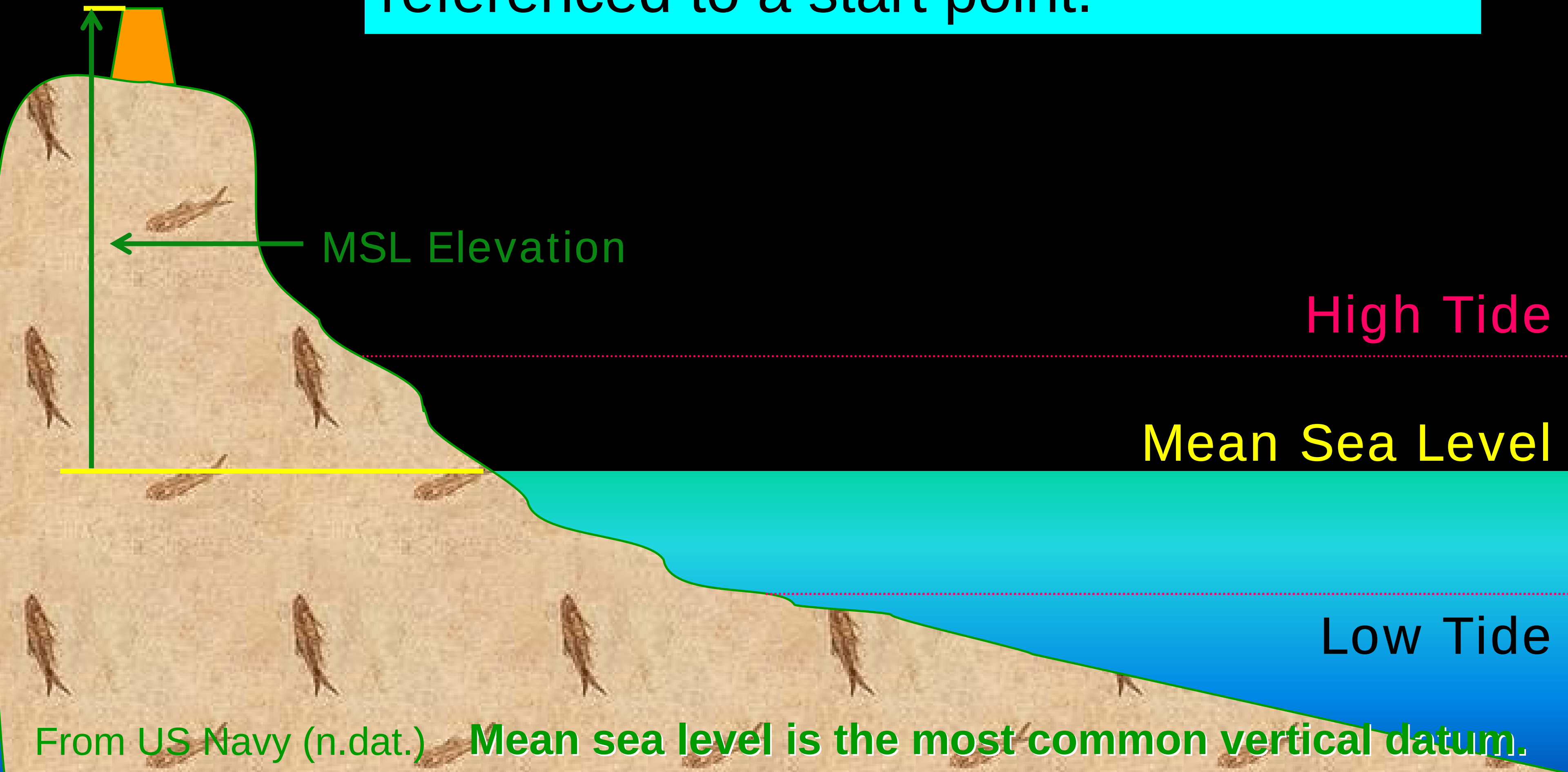
Australian Biodiversity Information Services

Differences between Datums

Datum from	Region or Location	Datum to	Difference
AGD66	Australia	AGD84	Max ± 0-5 m
AGD66/84	Australia	GDA94	Max ± 200 m
AGD66/84	Australia	WGS84	Max ± 200 m
GDA94	Australia	WGS84	Max ± <1 m
NAD 1983	North America	WGS84	Max ± <1 m
NAD27	North America	WGS84	Max ± 200 m
NAD 27	Contiguous USA	WGS84	Max ± 105 m
NAD 27	Aleutian Islands, Alaska	WGS84	Max ± 235 m
NAD 27	Hawaii	WGS 84	~ 500 m
TOKYO	Japan	WGS84	Max ± 750 m
ED-50	Europe	WGS84	Max ± 175 m
ARC-50	Africa	WGS84	Max ± 265 m
INDIAN 1975	Bangkok, Thailand	WGS84	~ 405 m
INDIAN 1956	Delhi, India	WGS84	~ 135 m
INDIAN 1956	Mumbai, India	WGS84	~ 120 m
HONG KONG 1973	Hong Kong	WGS84	~ 320 m
LUZON	Manila, The Philippines	WGS84	~ 225 m
TOKYO-KOREA	Seoul, South Korea	WGS84	~380 m
KERTAU 1948	Singapore	WGS84	~190 m

Vertical Datums

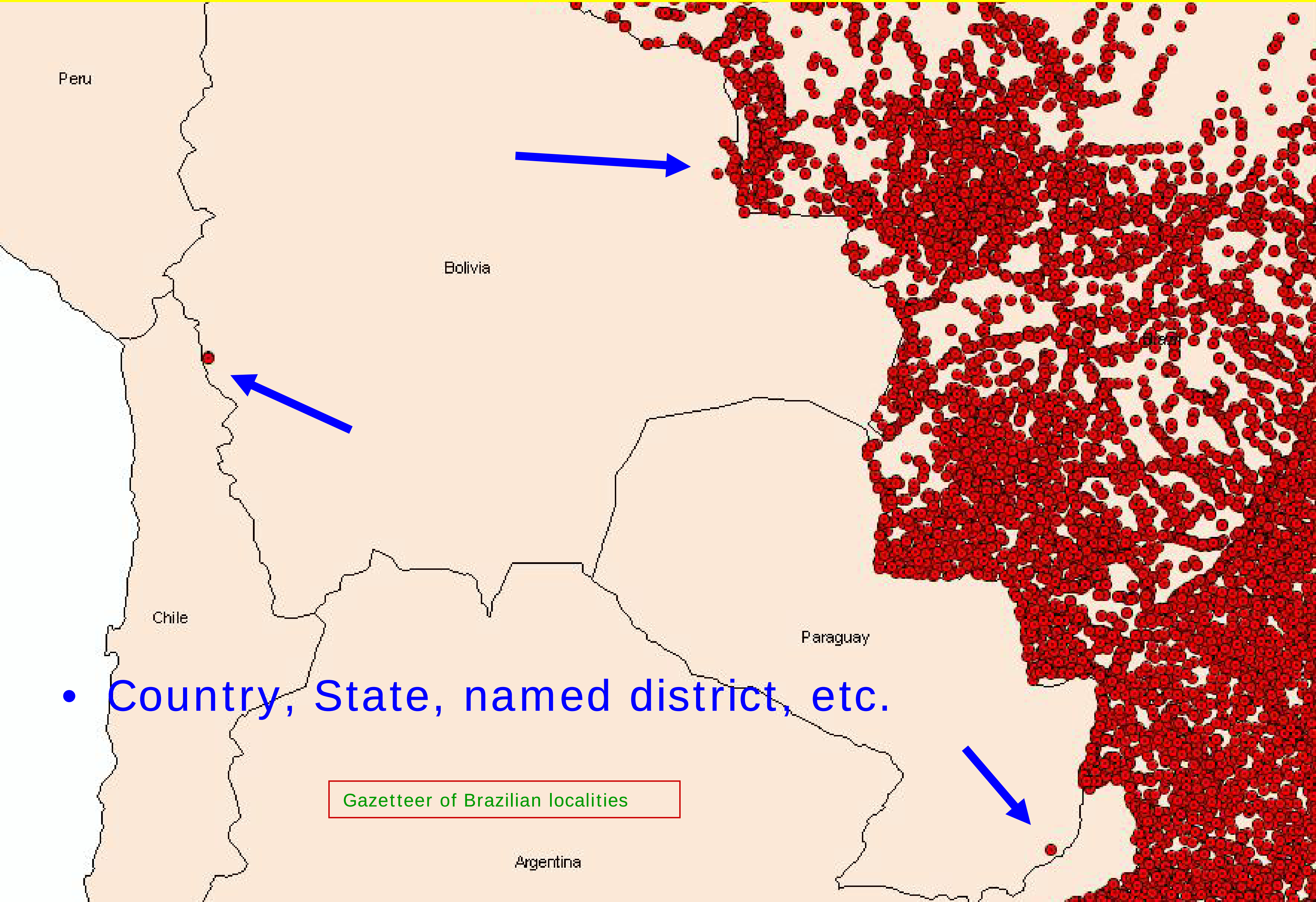
Like horizontal measurements, elevation only has meaning when referenced to a start point.



From US Navy (n.dat.)

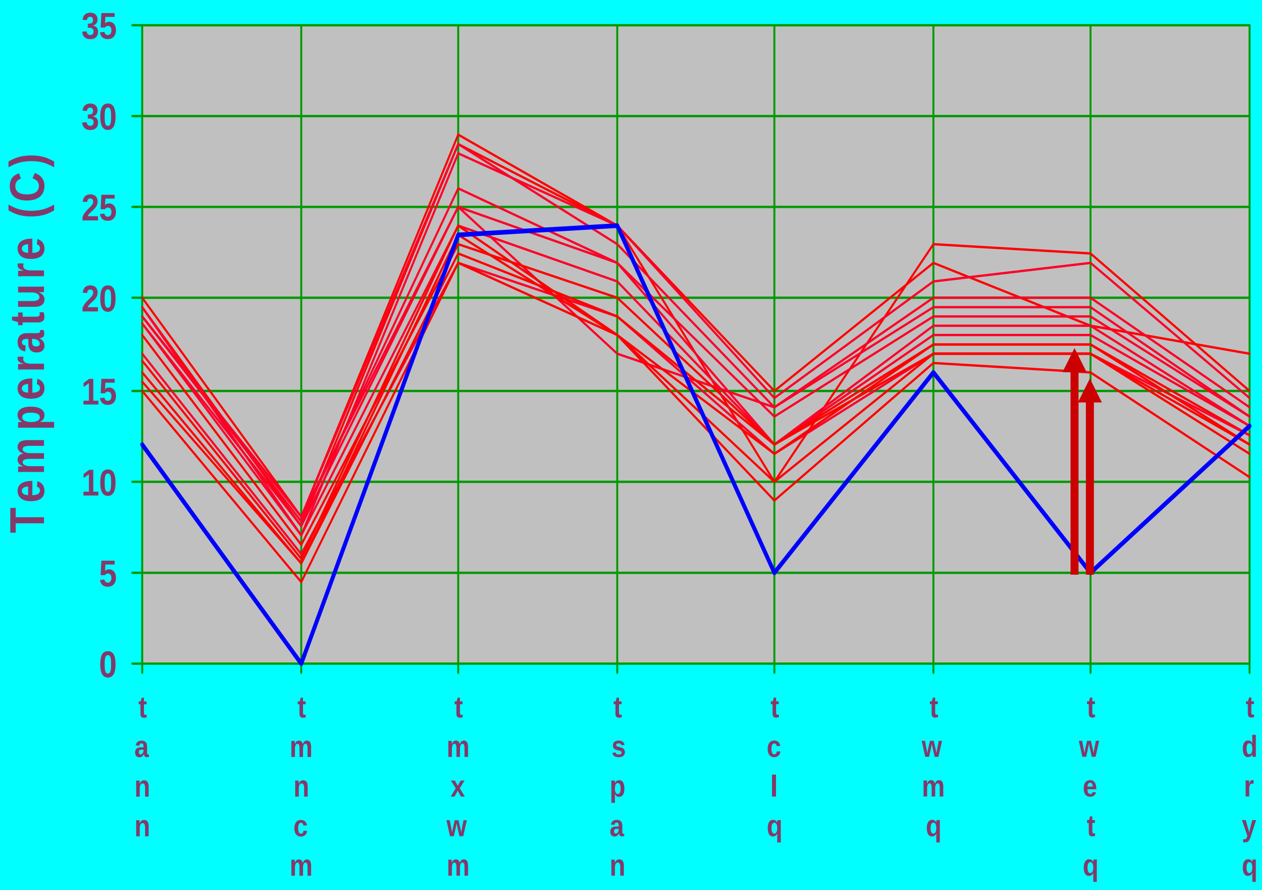
Mean sea level is the most common vertical datum.

Geographic outliers - GIS



- Country, State, named district, etc.

Using Climate to Identify Outliers



if $x < \bar{x}$

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(\bar{x} - x_{(i)}\right)$$

else

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(x_{(i+1)} - \bar{x}\right)$$

then

$$C = \frac{y_{(i)}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(y_{(i)} - \bar{y}\right)^2}{n-1}}}$$

Reverse Jack-knife

Acacia orites - 19 records -
9 Temperature parameters

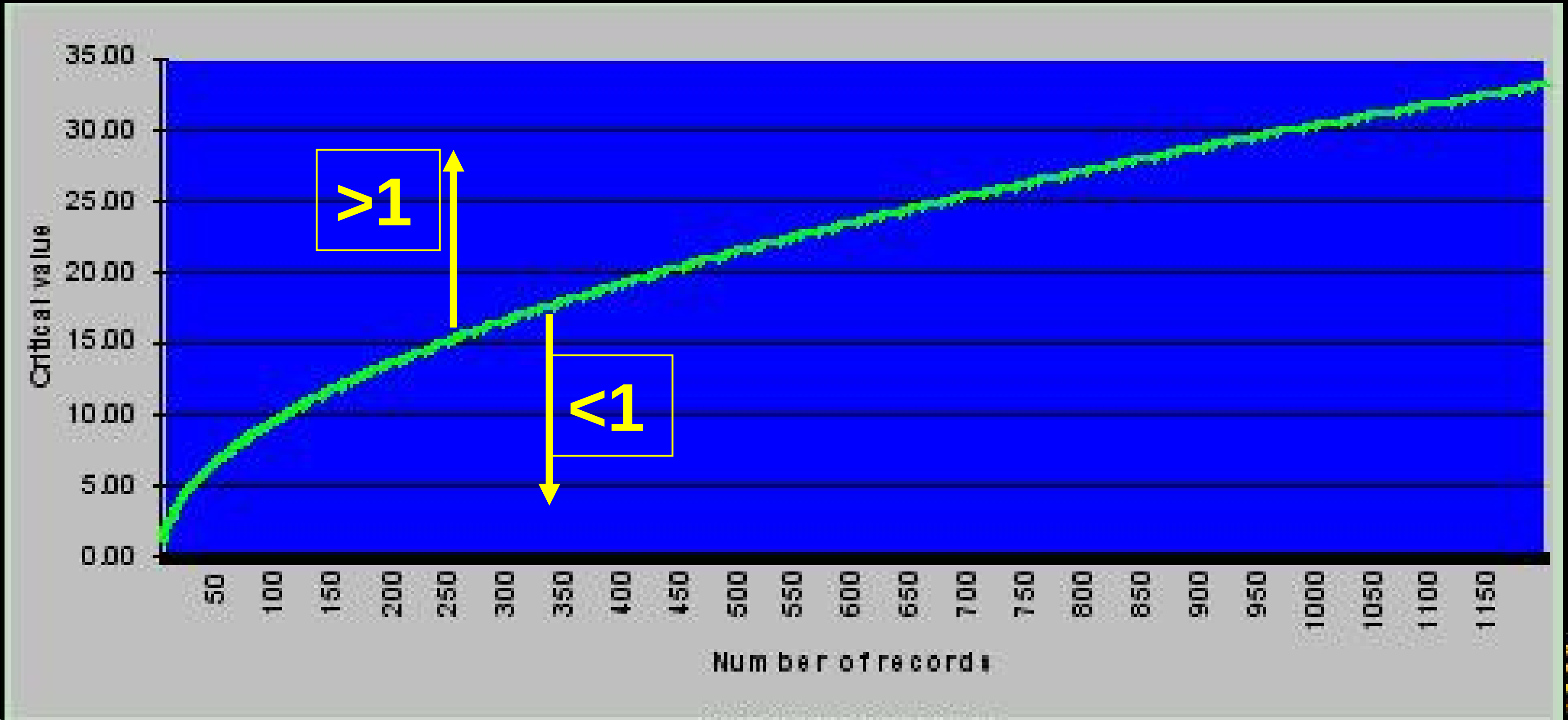
NB. Because the value of ‘C’ relates to it’s nearest point, successive values may be very small, so we ensure that if ‘x[i]’ is an outlier, then all points beyond are outliers too (even if they are clustered)

Concept of “*Outlierness*”

$T=((0.95(\sqrt{n})+0.2)\ X\ (Range/50))$
where ‘n’ is the number of records

“Outlierness” is the degree to which a record is an outlier

Outlierness = c[i]/ T



Araucaria bidwillii (Bunya Pine)

Outlierness (11/19)
11.0 (Isothermality)
10.0 (Precipitation Seasonality)
8.6 (Temp Seasonality)
8.6 (Precip Wettest Q)
8.3 (Precip Wettest Month)
7.7 (Precip Warmest Q)
5.7 (Ann Precip)
5.0 (Ann Temp Range)
4.5 (Mean Monthly Temp Range)
3.3 (Max Temp Warmest Month)
2.1 (Min Temp Coldest Month)
1.1 (Mean Temp Driest Q)

Outlierness (15/19)
12.0 (Isothermality)
10.9 (Precipitation Seasonality)
8.2 (Temp Seasonality)
6.3 (Mean Temp Driest Q)
5.8 (Min Temp Coldest Month)
5.4 (Mean Ann Temp)
5.2 (Mean Temp Coldest Q)
5.2 (Precipitation Wettest Q)
5.1 (Precip Wettest Month)
4.3 (Precipitation Warmest Q)
3.8 (Ann Temp Range)
2.9 (Mean Monthly Temp Range)
2.1 (Ann Precip)
1.8 (Precip Driest Month)
1.6 (Mean Temp Warmest Q)

After validation

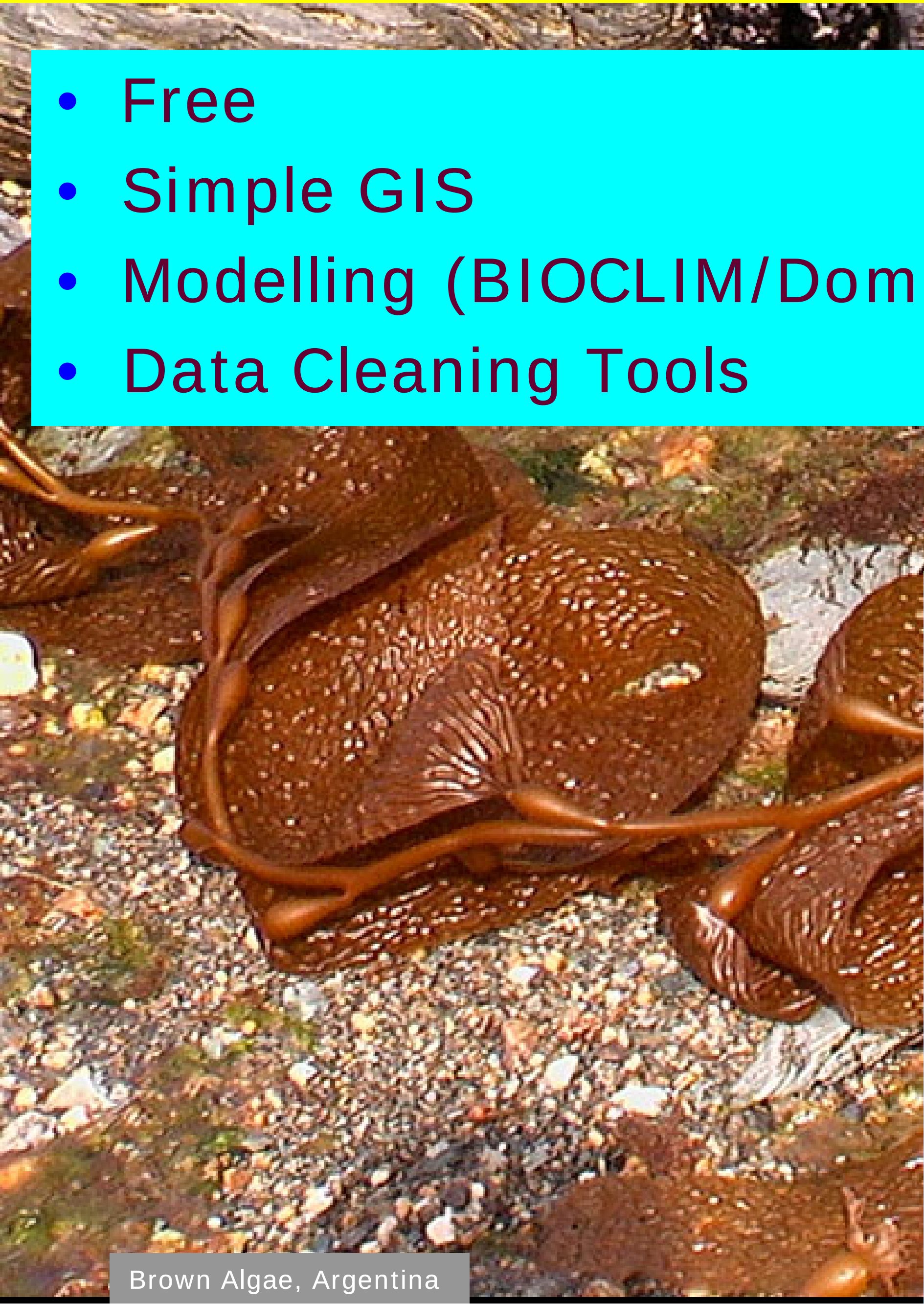
Araucaria bidwillii (Bunya Pine)

Outlierness (3/19)
3.6 (Max Temp Warmest Month)
3.4 (Mean Temp Warmest Q)
3.1 (Mean Temp Wettest Q)

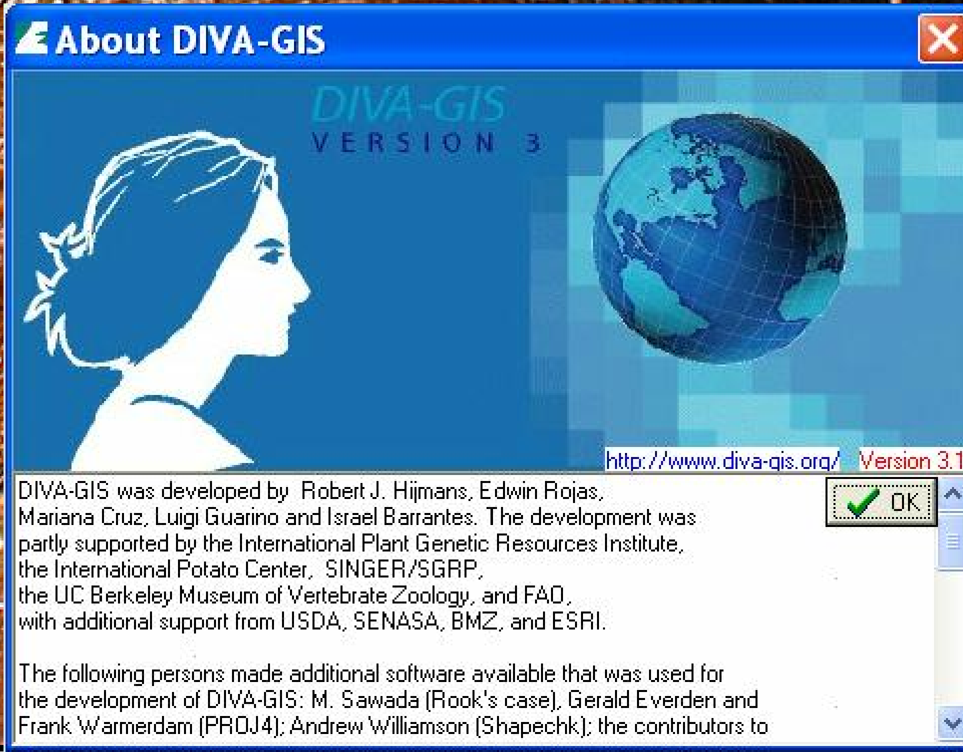
Outlierness (6/19)
6.0 (Precip Coldest Q)
4.0 (Precipitation Driest Q)
2.7 (Mean Ann Temp Range)
3.2 (Ann Precip)
2.7 (Mean Temp Driest Q)
1.2 (Isothermality)

Diva-GIS

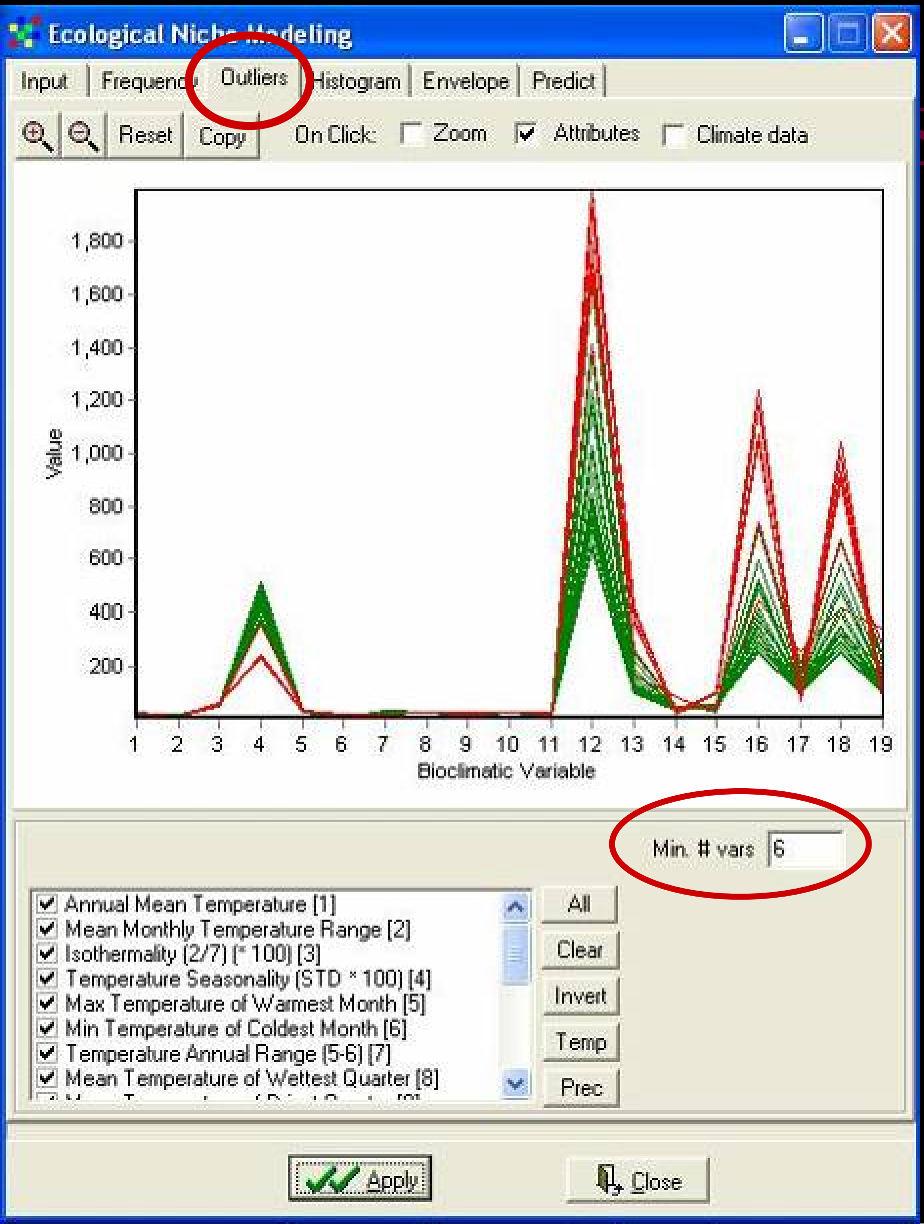
- Free
- Simple GIS
- Modelling (BIOCLIM/Domain)
- Data Cleaning Tools



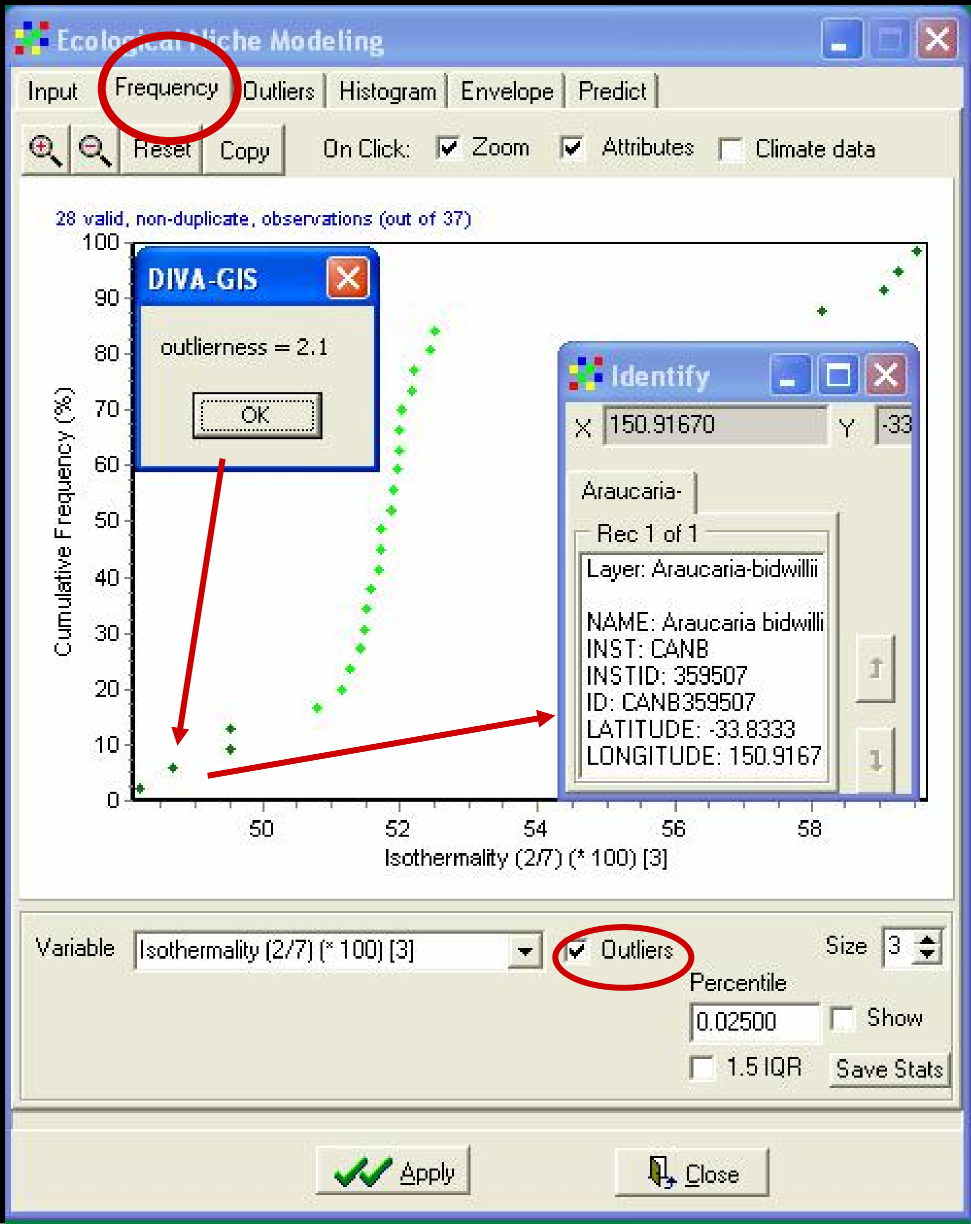
Brown Algae, Argentina



Diva-GIS

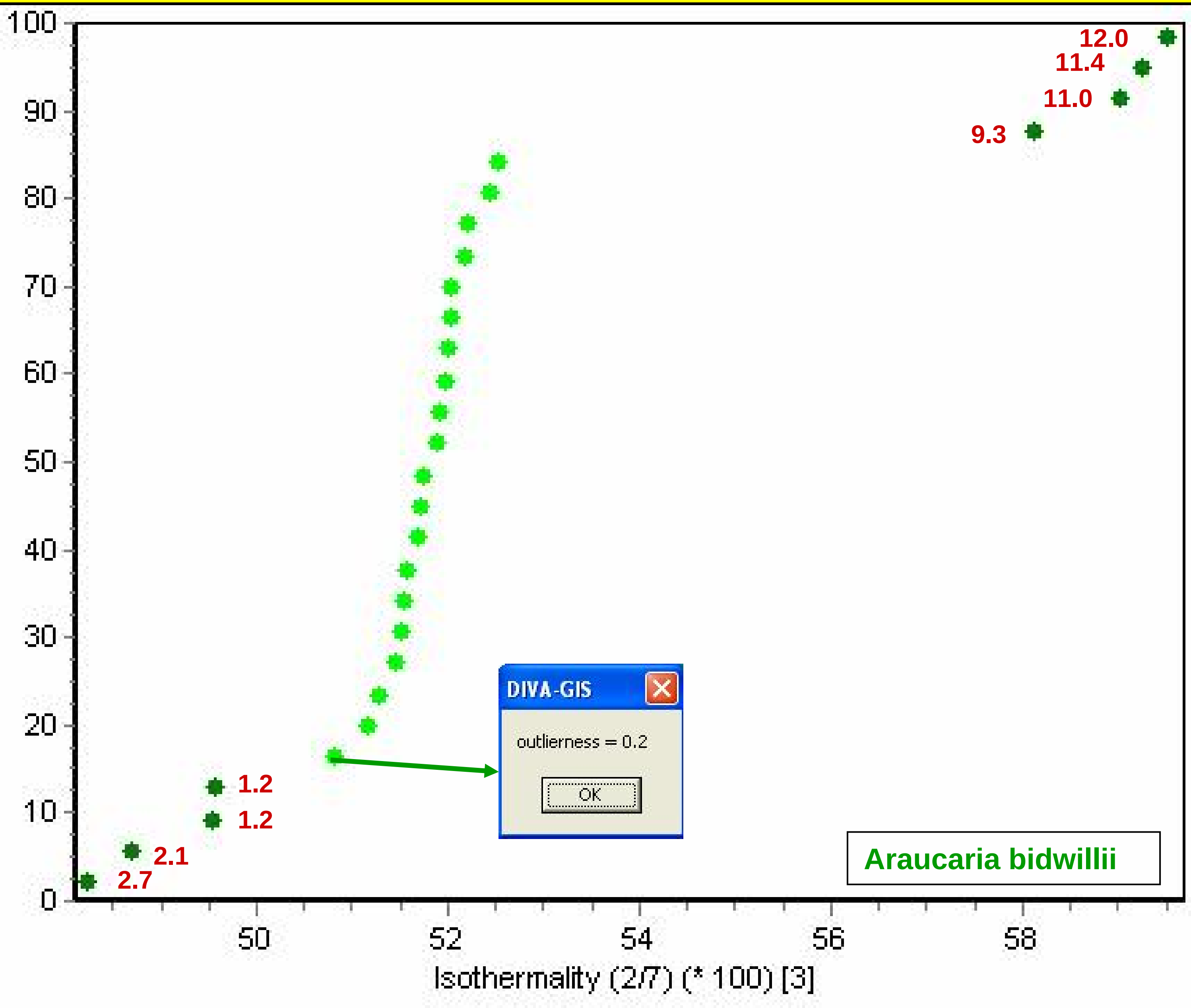


Ecological Niche Modeling
Input
Frequency
Outliers
Histogram
Envelope
Predict
Reset
Copy
On Click:
Zoom
Attributes
Climate data
Value
Bioclimatic Variable
Min. # vars
6
Annual Mean Temperature [1]
Mean Monthly Temperature Range [2]
Isothermality (2/7) (* 100) [3]
Temperature Seasonality (STD * 100) [4]
Max Temperature of Warmest Month [5]
Min Temperature of Coldest Month [6]
Temperature Annual Range (5-6) [7]
Mean Temperature of Wettest Quarter [8]
All
Clear
Invert
Temp
Prec
Apply
Close



Ecological Niche Modeling
Input
Frequency
Outliers
Histogram
Envelope
Predict
Reset
Copy
On Click:
Zoom
Attributes
Climate data
28 valid, non-duplicate, observations (out of 37)
Cumulative Frequency (%)
Isothermality (2/7) (* 100) [3]
DIVA-GIS
outlierness = 2.1
OK
Identify
X 150.91670
Y -33
Araucaria-
Rec 1 of 1
Layer: Araucaria-bidwillii
NAME: Araucaria bidwilli
INST: CANB
INSTID: 359507
ID: CANB359507
LATITUDE: -33.8333
LONGITUDE: 150.9167
Variable
Isothermality (2/7) (* 100) [3]
Outliers
Size
3
Percentile
0.02500
Show
1.5 IQR
Save Stats
Apply
Close

Outlierness values for Isothermality



Cumulative frequency curves can be used to demonstrate the "outlierness" concept

Australian Biodiversity Information Services

Errors in data

> *In general, error must not be treated as a potentially embarrassing inconvenience, because error provides a critical component in judging fitness for use.*
>
> Chrisman, 1991

> *Although most data gathering disciplines treat error as an embarrassing issue to be expunged, the error inherent in (spatial) data deserves closer attention and public understanding.*
>
> Chrisman, 1991

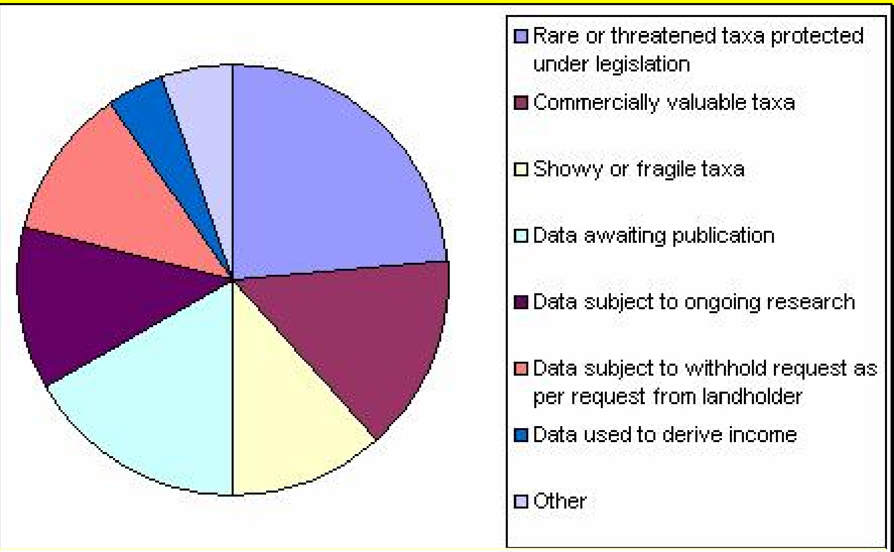
Australian Biodiversity Information Services

Challenges

– Dealing with sensitive data

Great Egret, Louisiana 2004

What constitutes sensitive taxa?



Summary of 93 responses to GBIF on-line Survey on Dealing with Sensitive Primary Species Occurrence Data

The issues:

- Institutions unknowingly making details of sensitive data available
- No universal global/regional list of sensitive taxa
- Duplicates of sensitive taxa from other countries

The Solutions?

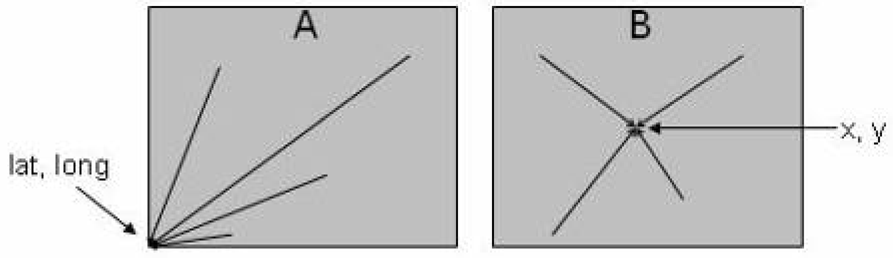
- Create an agreed list of sensitive taxa (by region)

Australian Biodiversity Information Services

Problems with generalizing

Institutions are generalizing in many different ways and not documenting what they are doing.

	Yes	Generalize (explain below)	Randomize (explain below)	Respondent Total
a. Remove altogether	**94% (16)**	18% (3)	0% (0)	**17**
b. Round to 1 minute	**86% (6)**	57% (4)	29% (2)	**7**
c. Round to 10 minutes	**83% (5)**	67% (4)	0% (0)	**6**
d. Round to 30 minutes	**60% (3)**	**60% (3)**	0% (0)	**5**
e. Round to 1 degree	25% (1)	**75% (3)**	0% (0)	**4**
f. Move to nearest named place	**67% (2)**	**67% (2)**	0% (0)	**3**
g. Report by geographic region	**65% (11)**	53% (9)	6% (1)	**17**
h. Report by bioregion	**86% (6)**	29% (2)	0% (0)	**7**
i. Report by standard grid	50% (7)	**71% (10)**	14% (2)	**14**
j. Report by map sheet (explain scale, etc. below)	**88% (7)**	62% (5)	12% (1)	**8**
k. Combination of >1 above (note which, below)	**100% (4)**	50% (2)	25% (1)	**4**
l. Some other method (explain below)	**89% (8)**	67% (6)	44% (4)	**9**
			Total Respondents	**46**
			(skipped this question)	56



Two generalization methods.

A. a geographic grid where all records are referenced to the bottom right-hand corner.

B. a metric grid where all records are referenced to the centroid.

Sociological Issues

- One case doesn't fit all
- Political issues
  - Endangered species (eg. Woller
  - National legislation
  - Piracy
  - Trade and Quaranti
- Privacy
  - Names of ... s, determiners
- Legal p... n
  - P... ns
  - ...servations in protected areas
  - Collections vis á vis permits
- uplicate Collections
  - Filtered Push

Solutions still to be worked out


Questions please?
Thank You!